

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Carol Peters Paul Clough
Fredric C. Gey Jussi Karlgren
Bernardo Magnini Douglas W. Oard
Maarten de Rijke Maximilian Stempfhuber (Eds.)

Evaluation of Multilingual and Multi-modal Information Retrieval

7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006
Alicante, Spain, September 20-22, 2006
Revised Selected Papers

Volume Editors

Carol Peters, ISTI-CNR, Pisa, Italy
E-mail: carol.peters@isti.cnr.it

Paul Clough, University of Sheffield, UK
E-mail: p.d.clough@sheffield.ac.uk

Fredric C. Gey, University of California, Berkeley, CA, USA
E-mail: gey@berkeley.edu

Jussi Karlgren, Swedish Institute of Computer Science, Kista, Sweden
E-mail: jussi@sics.se

Bernardo Magnini, FBK-irst, Trento, Italy
E-mail: magnini@itc.it

Douglas W. Oard, University of Maryland, College Park, MD, USA
E-mail: oard@glue.umd.edu

Maarten de Rijke, University of Amsterdam, Amsterdam, The Netherlands
E-mail: mdr@science.uva.nl

Maximilian Stempfhuber, GESIS-IZ, Bonn, Germany
E-mail: max.stempfhuber@gesis.org

Managing Editor
Danilo Giampiccolo, CELCT, Trento, Italy
E-mail: giampiccolo@celct.it

Library of Congress Control Number: 2007934753

CR Subject Classification (1998): H.3, I.2, H.4

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743

ISBN-10 3-540-74998-5 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-74998-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12161751 06/3180 5 4 3 2 1 0

Preface

The seventh campaign of the Cross Language Evaluation Forum (CLEF) for European languages was held from January to September 2006. There were eight evaluation tracks in CLEF 2006, designed to test the performance of a wide range of multilingual information access systems or system components. A total of 90 groups from all over the world submitted results for one or more of the evaluation tracks. Full details regarding the track design, the evaluation methodologies, and the results obtained by the participants can be found in the different sections of these proceedings. The results of the campaign were reported and discussed at the annual workshop, held in Alicante, Spain, 20-22 September, immediately following the 10th European Conference on Digital Libraries. The workshop was attended by approximately 130 researchers and system developers. In addition to presentations in plenary and parallel sessions, the poster session and breakout meetings gave participants a chance to discuss ideas and results in detail. An invited talk was given by Noriko Kando from the National Institute of Informatics, Tokyo on the NTCIR evaluation initiative for Asian languages. The final session focussed on technical transfer issues. Martin Braschler, from the University of Applied Sciences Winterthur, Switzerland, gave a talk on “What MLIA Applications Can Learn from Evaluation Campaigns” while Fredric Gey from the University of California, Berkeley, USA, summarised some of the main conclusions of the MLIA workshop at SIGIR 2006, where much of the discussion was concentrated on problems involved in building and marketing commercial MLIA systems. The workshop was preceded by two related events. On 19 September, the ImageCLEF group, together with the MUSCLE Network of Excellence, held a joint meeting on “Usage-Oriented Multimedia Information Retrieval Evaluation”. On the morning of 20 September, before the official beginning of the workshop, members of the question answering group, coordinated by Fernando Llopis, University of Alicante, organised an exercise designed to test the ability of question answering systems to respond within a time constraint. This was the first time that an activity of this type had been held at CLEF; it was a great success and aroused much interest. It is our intention to repeat this experience at CLEF 2007. The presentations given at the workshop can be found on the CLEF website at: www.clef-campaign.org. These post-campaign proceedings represent extended and revised versions of the initial working notes distributed at the workshop. All papers have been subjected to a reviewing procedure. The volume has been prepared with the assistance of the Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy, under the coordination of Danilo Giampiccolo. The support of CELCT is gratefully acknowledged. We should also like to thank all our reviewers for their careful refereeing. CLEF 2006 was an activity of the DELOS Network of Excellence for

Digital Libraries, within the framework of the Information Society Technologies programme of the European Commission.

July 2007

Carol Peters
Douglas W. Oard
Paul Clough
Fredric C. Gey
Max Stempfhuber
Maarten de Rijke
Jussi Karlgren
Bernardo Magnini

Reviewers

The Editors express their gratitude to the colleagues listed below for their assistance in reviewing the papers in this volume:

- Christelle Ayache, ELDA/ELRA, Evaluations and Language Resources Distribution Agency, Paris, France
- Krisztian Balog, University of Amsterdam, The Netherlands
- Thomas Deselaers, Lehrstuhl für Informatik 6, Aachen University of Technology (RWTH), Germany
- Giorgio Di Nunzio, Dept. of Information Engineering, University of Padua, Italy
- Nicola Ferro, Dept. of Information Engineering, University of Padua, Italy
- Cam Fordyce, CELCT, Center for the Evaluation of Language and Communication Technologies, Trento, Italy
- Pamela Forner, CELCT, Center for the Evaluation of Language and Communication Technologies, Trento, Italy
- Danilo Giampiccolo, CELCT, Center for the Evaluation of Language and Communication Technologies, Trento, Italy
- Michael Grubinger, School of Computer Science and Mathematics, Victoria University, Melbourne, Australia
- Allan Hanbury, Vienna University of Technology, Austria
- William Hersh, Dept. of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, USA
- Valentin Jijkoun, University of Amsterdam, The Netherlands
- Gareth J.F. Jones, Dublin City University, Ireland
- Jayashree Kalpathy-Cramer, Dept. of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, USA
- Jaap Kamps, Archive and Documentation Studies, University of Amsterdam, The Netherlands
- Thomas M. Lehmann, Dept. of Medical Informatics, Aachen University of Technology (RWTH), Germany
- Thomas Mandl, Information Science, University of Hildesheim, Germany
- Andrés Montoyo, University of Alicante, Spain
- Henning Müller, University and University Hospitals of Geneva, Switzerland
- Petya Osenova, BulTreeBank Project, CLPP, Bulgarian Academy of Sciences, Sofia, Bulgaria
- Paulo Rocha, Linguatca, Sintef ICT, Oslo, Norway, and Braga, Lisbon & Porto, Portugal
- Bogdan Sacaleanu, DFKI, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany
- Diana Santos, Linguatca Sintef, Oslo, Norway
- Richard Sutcliffe, University of Limerick, Ireland

CLEF 2006 Coordination

CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa. The following Institutions contributed to the organisation of the different tracks of the CLEF 2006 campaign:

- Center for the Evaluation of Language and Communication Technologies (CELCT), Trento, Italy
- Centro per la Ricerca Scientifica e Tecnologica, ITC, Trento, Italy
- College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, USA
- Department of Computer Science, University of Helsinki, Finland
- Department of Computer Science, University of Indonesia
- Department of Computer Science, RWTH Aachen University, Germany
- Department of Computer Science and Information Systems, University of Limerick, Ireland
- Department of Computer Science and Information Engineering, National University of Taiwan
- Department of Information Engineering, University of Padua, Italy
- Department of Information Science, University of Hildesheim, Germany
- Department of Information Studies, University of Sheffield, UK
- Department of Medical Informatics, RWTH Aachen University, Germany
- Evaluation and Language Resources Distribution Agency (ELDA), France
- Research Center for Artificial Intelligence, DFKI, Saarbrücken, Germany
- Information and Language Processing Systems, University of Amsterdam, The Netherlands
- Informationszentrum Sozialwissenschaften, Bonn, Germany
- Institute for Information Technology, Hyderabad, India
- Lenguajes y Sistemás Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
- LINGUECA, Sintef, Oslo, Norway
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Bulgaria
- National Institute of Standards and Technology, Gaithersburg MD, USA
- Oregon Health and Science University, USA
- Research Computing Center of Moscow State University, Russia
- Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- School of Computer Science, Charles University, Prague, Czech Republic
- School of Computer Science and Mathematics, Victoria University, Australia
- School of Computing, Dublin City University, Ireland
- UC Data Archive and School of Information Management and Systems, UC Berkeley, USA
- University "Alexandru Ioan Cuza", IASI, Romania
- University and University Hospitals of Geneva, Switzerland

CLEF 2006 Steering Committee

- Maristella Agosti, University of Padua, Italy
- Martin Braschler, Zurich University of Applied Sciences Winterthur, Switzerland
- Amedeo Cappelletti, ISTI-CNR & CELCT, Italy
- Hsin-Hsi Chen, National Taiwan University, Taipei, Taiwan
- Khalid Choukri, Evaluations and Language Resources Distribution Agency, Paris, France
- Paul Clough, University of Sheffield, UK
- Thomas Deselaers, RWTH Aachen University, Germany
- David A. Evans, Clairvoyance Corporation, USA
- Marcello Federico, ITC-irst, Trento, Italy
- Christian Fluhr, CEA-LIST, Fontenay-aux-Roses, France
- Norbert Fuhr, University of Duisburg, Germany
- Frederic C. Gey, U.C. Berkeley, USA
- Julio Gonzalo, LSI-UNED, Madrid, Spain
- Donna Harman, National Institute of Standards and Technology, USA
- Gareth Jones, Dublin City University, Ireland
- Franciska de Jong, University of Twente, The Netherlands
- Noriko Kando, National Institute of Informatics, Tokyo, Japan
- Jussi Karlgren, Swedish Institute of Computer Science, Sweden
- Michael Kluck, Informationszentrum Sozialwissenschaften Bonn, Germany
- Natalia Loukachevitch, Moscow State University, Russia
- Bernardo Magnini, ITC-irst, Trento, Italy
- Paul McNamee, Johns Hopkins University, USA
- Henning Müller, University & Hospitals of Geneva, Switzerland
- Douglas W. Oard, University of Maryland, USA
- Maarten de Rijke, University of Amsterdam, The Netherlands
- Diana Santos, Linguatca, Sintef, Oslo, Norway
- Jacques Savoy, University of Neuchâtel, Switzerland
- Peter Schauble, Eurospider Information Technologies, Switzerland
- Max Stempfhuber, Informationszentrum Sozialwissenschaften Bonn, Germany
- Richard Sutcliffe, University of Limerick, Ireland
- Hans Uszkoreit, German Research Center for Artificial Intelligence (DFKI), Germany
- Felisa Verdejo, LSI-UNED, Madrid, Spain
- José Luis Vicedo, University of Alicante, Spain
- Ellen Voorhees, National Institute of Standards and Technology, USA
- Christa Womser-Hacker, University of Hildesheim, Germany

Table of Contents

Introduction

What Happened in CLEF 2006	1
<i>Carol Peters</i>	
Scientific Data of an Evaluation Campaign: Do We Properly Deal with Them?	11
<i>Maristella Agosti, Giorgio Maria Di Nunzio, and Nicola Ferro</i>	

Part I: Multilingual Textual Document Retrieval (Ad Hoc)

CLEF 2006: Ad Hoc Track Overview	21
<i>Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters</i>	

Cross-Language

Hindi, Telugu, Oromo, English CLIR Evaluation	35
<i>Prasad Pingali, Kula Kekeba Tune, and Vasudeva Varma</i>	
Amharic-English Information Retrieval	43
<i>Atelach Alemu Argaw and Lars Asker</i>	
The University of Lisbon at CLEF 2006 Ad-Hoc Task	51
<i>Nuno Cardoso, Mário J. Silva, and Bruno Martins</i>	
Query and Document Translation for English-Indonesian Cross Language IR	57
<i>Herika Hayurani, Syandra Sari, and Mirna Adriani</i>	

Monolingual

Passage Retrieval vs. Document Retrieval in the CLEF 2006 Ad Hoc Monolingual Tasks with the IR-n System	62
<i>Elisa Noguera and Fernando Llopis</i>	
The PUCRS NLP-Group Participation in CLEF2006: Information Retrieval Based on Linguistic Resources	66
<i>Marco Gonzalez and Vera Lúcia Strube de Lima</i>	

NLP-Driven Constructive Learning for Filtering an IR Document Stream	74
<i>João Marcelo Azevedo Arcoverde and Maria das Graças Volpe Nunes</i>	
ENSM-SE at CLEF 2006 : Fuzzy Proximity Method with an Adhoc Influence Function	83
<i>Annabelle Mercier and Michel Beigbeder</i>	
A Study on the Use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval	91
<i>Viviane Moreira Orengo, Luciana S. Buriol, and Alexandre Ramos Coelho</i>	
Benefits of Resource-Based Stemming in Hungarian Information Retrieval	99
<i>Péter Halácsy and Viktor Trón</i>	
Statistical vs. Rule-Based Stemming for Monolingual French Retrieval	107
<i>Prasenjit Majumder, Mandar Mitra, and Kalyankumar Datta</i>	

Robust and More

A First Approach to CLIR Using Character <i>N</i> -Grams Alignment	111
<i>Jesús Vilares, Michael P. Oakes, and John I. Tait</i>	
SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: Query Expansion Using the Google Search Engine	119
<i>Fernando Martínez-Santiago, Arturo Montejo-Ráez, Miguel Á. García-Cumbreras, and L. Alfonso Ureña-López</i>	
Robust Ad-Hoc Retrieval Experiments with French and English at the University of Hildesheim	127
<i>Thomas Mandl, René Hackl, and Christa Womser-Hacker</i>	
Comparing the Robustness of Expansion Techniques and Retrieval Measures	129
<i>Stephen Tomlinson</i>	
Experiments with Monolingual, Bilingual, and Robust Retrieval	137
<i>Jacques Savoy and Samir Abdou</i>	
Local Query Expansion Using Terms Windows for Robust Retrieval	145
<i>Angel F. Zazo, Jose L. Alonso Berrocal, and Carlos G. Figuerola</i>	
Dublin City University at CLEF 2006: Robust Cross Language Track ...	153
<i>Adenike M. Lam-Adesina and Gareth J.F. Jones</i>	

JHU/APL Ad Hoc Experiments at CLEF 2006	157
<i>Paul McNamee</i>	

Part II: Domain-Specific Information Retrieval (Domain-Specific)

The Domain-Specific Track at CLEF 2006: Overview of Approaches, Results and Assessment	163
<i>Maximilian Stempfhuber and Stefan Baerisch</i>	
Reranking Documents with Antagonistic Terms	170
<i>Johannes Leveling</i>	
Domain Specific Retrieval: Back to Basics	174
<i>Ray R. Larson</i>	
Monolingual Retrieval Experiments with a Domain-Specific Document Corpus at the Chemnitz University of Technology	178
<i>Jens Kürsten and Maximilian Eibl</i>	

Part III: Interactive Cross-Language Information Retrieval (i-CLEF)

iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository	186
<i>Jussi Karlgren, Julio Gonzalo, and Paul Clough</i>	
Are Users Willing to Search Cross-Language? An Experiment with the Flickr Image Sharing Repository	195
<i>Javier Artilles, Julio Gonzalo, Fernando López-Ostenero, and V́ctor Peinado</i>	
Providing Multilingual Access to FLICKR for Arabic Users	205
<i>Paul Clough, Azzah Al-Maskari, and Kareem Darwish</i>	
Trusting the Results in Cross-Lingual Keyword-Based Image Retrieval	217
<i>Jussi Karlgren and Fredrik Olsson</i>	

Part IV: Multiple Language Question Answering (QA@CLEF)

Overview of the CLEF 2006 Multilingual Question Answering Track	223
<i>Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe</i>	

Overview of the Answer Validation Exercise 2006	257
<i>Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo</i>	
Overview of the WiQA Task at CLEF 2006	265
<i>Valentin Jijkoun and Maarten de Rijke</i>	

Main Task: Mono- and Bilingual QA

Re-ranking Passages with LSA in a Question Answering System	275
<i>David Tomás and José L. Vicedo</i>	
Question Types Specification for the Use of Specialized Patterns in Prodicos System	280
<i>E. Desmontils, C. Jacquin, and L. Monceaux</i>	
Answer Translation: An Alternative Approach to Cross-Lingual Question Answering	290
<i>Johan Bos and Malvina Nissim</i>	
Priberam's Question Answering System in a Cross-Language Environment	300
<i>Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, Cláudia Pinto, and Daniel Vidal</i>	
LCC's PowerAnswer at QA@CLEF 2006	310
<i>Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Jonathan Clark, and Dan Moldovan</i>	
Using Syntactic Knowledge for QA	318
<i>Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann</i>	
A Cross-Lingual German-English Framework for Open-Domain Question Answering	328
<i>Bogdan Sacaleanu and Günter Neumann</i>	
Cross Lingual Question Answering Using QRISTAL for CLEF 2006	339
<i>Dominique Laurent, Patrick Séguéla, and Sophie Nègre</i>	
CLEF2006 Question Answering Experiments at Tokyo Institute of Technology	351
<i>E.W.D. Whittaker, J.R. Novak, P. Chatain, P.R. Dixon, M.H. Heie, and S. Furui</i>	
Quartz: A Question Answering System for Dutch	362
<i>David Ahn, Valentin Jijkoun, Joris van Rantwijk, Maarten de Rijke, and Erik Tjong Kim Sang</i>	

Experiments on Applying a Text Summarization System for Question Answering	372
<i>Pedro Paulo Balage Filho, Vinícius Rodrigues de Uzêda, Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes</i>	
N-Gram vs. Keyword-Based Passage Retrieval for Question Answering	377
<i>Davide Buscaldi, José Manuel Gomez, Paolo Rosso, and Emilio Sanchis</i>	
Cross-Lingual Romanian to English Question Answering at CLEF 2006	385
<i>Georgiana Puşcaşu, Adrian Iftene, Ionuţ Pistol, Diana Trandabăţ, Dan Tufiş, Alin Ceaşu, Dan Ştefănescu, Radu Ion, Iustin Dornescu, Alex Moruz, and Dan Cristea</i>	
Finding Answers in the Œdipe System by Extracting and Applying Linguistic Patterns	395
<i>Romaric Besançon, Mehdi Embarek, and Olivier Ferret</i>	
Question Answering Beyond CLEF Document Collections	405
<i>Luís Costa</i>	
Using Machine Learning and Text Mining in Question Answering	415
<i>Antonio Juárez-González, Alberto Téllez-Valero, Claudia Denicia-Carral, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda</i>	
Applying Dependency Trees and Term Density for Answer Selection Reinforcement	424
<i>Manuel Pérez-Coutiño, Manuel Montes-y-Gómez, Aurelio López-López, Luis Villaseñor-Pineda, and Aarón Pancardo-Rodríguez</i>	
Interpretation and Normalization of Temporal Expressions for Question Answering	432
<i>Sven Hartrumpf and Johannes Leveling</i>	
Relevance Measures for Question Answering, The LIA at QA@CLEF-2006	440
<i>Laurent Gillard, Laurianne Sitbon, Eric Blaudez, Patrice Bellot, and Marc El-Bèze</i>	
Monolingual and Cross-Lingual QA Using AliQAn and BRILI Systems for CLEF 2006	450
<i>S. Ferrández, P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alvarado, E. Noguera, and F. Llopis</i>	

The Bilingual System MUSCLEF at QA@CLEF 2006	454
<i>Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat Michael Bagur, and Kevin Séjourné</i>	
MIRACLE Experiments in QA@CLEF 2006 in Spanish: Main Task, Real-Time QA and Exploratory QA Using Wikipedia (WiQA)	463
<i>César de Pablo-Sánchez, Ana González-Ledesma, Antonio Moreno-Sandoval, and Maria Teresa Vicente-Díez</i>	
A First Step to Address Biography Generation as an Iterative QA Task	473
<i>Luis Sarmento</i>	

Answer Validation Exercise (AVE)

The Effect of Entity Recognition on Answer Validation	483
<i>Álvaro Rodrigo, Anselmo Peñas, Jesús Herrera, and Felisa Verdejo</i>	
A Knowledge-Based Textual Entailment Approach Applied to the AVE Task	490
<i>Ó. Ferrández, R.M. Terol, R. Muñoz, P. Martínez-Barco, and M. Palomar</i>	
Automatic Answer Validation Using COGEX	494
<i>Marta Tatu, Brandon Iles, and Dan Moldovan</i>	
Paraphrase Substitution for Recognizing Textual Entailment	502
<i>Wauter Bosma and Chris Callison-Burch</i>	
Experimenting a “General Purpose” Textual Entailment Learner in AVE	510
<i>Fabio Massimo Zanzotto and Alessandro Moschitti</i>	
Answer Validation Through Robust Logical Inference	518
<i>Ingo Glöckner</i>	
University of Alicante at QA@CLEF2006: Answer Validation Exercise	522
<i>Zornitsa Kozareva, Sonia Vázquez, and Andrés Montoyo</i>	
Towards Entailment-Based Question Answering: ITC-irst at CLEF 2006	526
<i>Milen Kouylekov, Matteo Negri, Bernardo Magnini, and Bonaventura Coppola</i>	

Question Answering Using Wikipedia (WiQA)

Link-Based vs. Content-Based Retrieval for Question Answering Using Wikipedia	537
<i>Sisay Fissaha Adafre, Valentin Jijkoun, and Maarten de Rijke</i>	

Identifying Novel Information Using Latent Semantic Analysis in the WiQA Task at CLEF 2006	541
<i>Richard F.E. Sutcliffe, Josef Steinberger, Udo Kruschwitz, Mijail Alexandrov-Kabadjov, and Massimo Poesio</i>	
A Bag-of-Words Based Ranking Method for the Wikipedia Question Answering Task	550
<i>Davide Buscaldi and Paolo Rosso</i>	
University of Alicante at WiQA 2006	554
<i>Antonio Toral Ruiz, Georgiana Puşcaşu, Lorenza Moreno Monteagudo, Rubén Izquierdo Beviá, and Estela Saquete Boró</i>	
A High Precision Information Retrieval Method for WiQA	561
<i>Constantin Orăsan and Georgiana Puşcaşu</i>	
QoLA: Fostering Collaboration Within QA	569
<i>Diana Santos and Luís Costa</i>	

Part V: Cross-Language Retrieval in Image Collections (ImageCLEF)

Overviews

Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks	579
<i>Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller</i>	
Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks	595
<i>Henning Müller, Thomas Deselaers, Thomas Deserno, Paul Clough, Eugene Kim, and William Hersh</i>	

ImageCLEFphoto

Text Retrieval and Blind Feedback for the ImageCLEFphoto Task	609
<i>Ray R. Larson</i>	
Expanding Queries Through Word Sense Disambiguation	613
<i>J.L. Martínez-Fernández, Ana M. García-Serrano, Julio Villena Román, and Paloma Martínez</i>	
Using Visual Linkages for Multilingual Image Retrieval	617
<i>Masashi Inoue</i>	

Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval	625
<i>Yih-Chen Chang and Hsin-Hsi Chen</i>	

Dublin City University at CLEF 2006: Experiments for the ImageCLEF Photo Collection Standard Ad Hoc Task	633
<i>Kieran McDonald and Gareth J.F. Jones</i>	

ImageCLEFmed

Image Classification with a Frequency-Based Information Retrieval Scheme for ImageCLEFmed 2006	638
<i>Henning Müller, Tobias Gass, and Antoine Geissbuhler</i>	

Grayscale Radiograph Annotation Using Local Relational Features	644
<i>Lokesh Setia, Alexandra Teynor, Alaa Halawani, and Hans Burkhardt</i>	

MorphoSaurus in ImageCLEF 2006: The Effect of Subwords on Biomedical IR	652
<i>Philipp Daumke, Jan Paetzold, and Kornel Marko</i>	

Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006	660
<i>William Hersh, Jayashree Kalpathy-Cramer, and Jeffery Jensen</i>	

MedIC at ImageCLEF 2006: Automatic Image Categorization and Annotation Using Combined Visual Representations	670
<i>Filip Florea, Alexandrina Rogozan, Eugen Barbu, Abdelaziz Bensrhair, and Stefan Darmoni</i>	

Medical Image Annotation and Retrieval Using Visual Features	678
<i>Jing Liu, Yang Hu, Mingjing Li, Songde Ma, and Wei-ying Ma</i>	

Baseline Results for the ImageCLEF 2006 Medical Automatic Annotation Task	686
<i>Mark O. Güld, Christian Thies, Benedikt Fischer, and Thomas M. Deserno</i>	

A Refined SVM Applied in Medical Image Annotation	690
<i>Bo Qiu</i>	

Inter-media Concept-Based Medical Image Indexing and Retrieval with UMLS at IPAL	694
<i>Caroline Lacoste, Jean-Pierre Chevallet, Joo-Hwee Lim, Diem Thi Hoang Le, Wei Xiong, Daniel Racoceanu, Roxana Teodorescu, and Nicolas Vuillenemot</i>	

UB at ImageCLEFmed 2006	702
<i>Miguel E. Ruiz</i>	

ImageCLEFphoto and med

Translation by Text Categorisation: Medical Image Retrieval in ImageCLEFmed 2006.....	706
<i>Julien Gobeill, Henning Müller, and Patrick Ruch</i>	
Using Information Gain to Improve the ImageCLEF 2006 Collection....	711
<i>M.C. Díaz-Galiano, M.Á. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L. Alfonso Ureña-López</i>	
CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections	715
<i>M.M. Rahman, V. Sood, B.C. Desai, and P. Bhattacharya</i>	
Image Retrieval and Annotation Using Maximum Entropy	725
<i>Thomas Deselaers, Tobias Weyand, and Hermann Ney</i>	
Inter-media Pseudo-relevance Feedback Application to ImageCLEF 2006 Photo Retrieval	735
<i>Nicolas Maillot, Jean-Pierre Chevallet, and Joo Hwee Lim</i>	
ImageCLEF 2006 Experiments at the Chemnitz Technical University ...	739
<i>Thomas Wilhelm and Maximilian Eibl</i>	

Part VI: Cross-Language Speech Retrieval (CLSR)

Overview of the CLEF-2006 Cross-Language Speech Retrieval Track....	744
<i>Douglas W. Oard, Jianqiang Wang, Gareth J.F. Jones, Ryen W. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran</i>	
Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006	759
<i>Pavel Ircing and Luděk Müller</i>	
Applying Logic Forms and Statistical Methods to CL-SR Performance.....	766
<i>Rafael M. Terol, Patricio Martinez-Barco, and Manuel Palomar</i>	
XML Information Retrieval from Spoken Word Archives	770
<i>Robin Aly, Djoerd Hiemstra, Roeland Ordelman, Laurens van der Werff, and Franciska de Jong</i>	
Experiments for the Cross Language Speech Retrieval Task at CLEF 2006	778
<i>Muath Alzghool and Diana Inkpen</i>	
CLEF-2006 CL-SR at Maryland: English and Czech	786
<i>Jianqiang Wang and Douglas W. Oard</i>	

Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments	794
<i>Gareth J.F. Jones, Ke Zhang, and Adenike M. Lam-Adesina</i>	

Part VII: Multilingual Web Track (WebCLEF)

Overview of WebCLEF 2006	803
<i>Krisztian Balog, Leif Azzopardi, Jaap Kamps, and Maarten de Rijke</i>	
Improving Web Pages Retrieval Using Combined Fields	820
<i>Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez, and Emilio Rodríguez</i>	
A Penalisation-Based Ranking Approach for the Mixed Monolingual Task of WebCLEF 2006	826
<i>David Pinto, Paolo Rosso, and Ernesto Jiménez</i>	
Index Combinations and Query Reformulations for Mixed Monolingual Web Retrieval	830
<i>Krisztian Balog and Maarten de Rijke</i>	
Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for WebCLEF 2006 at the University of Hildesheim	834
<i>Ben Heuwing, Thomas Mandl, and Robert Strötgen</i>	
Vocabulary Reduction and Text Enrichment at WebCLEF	838
<i>Franco Rojas, Héctor Jiménez-Salazar, and David Pinto</i>	
Experiments with the 4 Query Sets of WebCLEF 2006	844
<i>Stephen Tomlinson</i>	
Applying Relevance Feedback for Retrieving Web-Page Retrieval	848
<i>Syntia Wijaya, Bimo Widhi, Tommy Khoerniawan, and Mirna Adriani</i>	

Part VIII: Cross-Language Geographical Retrieval (GeoCLEF)

GeoCLEF 2006: The CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview	852
<i>Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio M. Di Nunzio, and Nicola Ferro</i>	
MIRACLE's Ad-Hoc and Geographical IR Approaches for CLEF 2006	877
<i>José M. Goñi-Menoyo, José C. González-Cristóbal, Sara Lana-Serrano, and Ángel Martínez-González</i>	

GIR Experimentation	881
<i>Andogah Geoffrey</i>	
GIR with Geographic Query Expansion	889
<i>Antonio Toral, Oscar Ferrández, Elisa Noguera, Zornitsa Kozareva, Andrés Montoyo, and Rafael Muñoz</i>	
Monolingual and Bilingual Experiments in GeoCLEF2006	893
<i>Rocio Guillén</i>	
Experiments on the Exclusion of Metonymic Location Names from GIR	901
<i>Johannes Leveling and Dirk Veiel</i>	
The University of New South Wales at GeoCLEF 2006	905
<i>You-Heng Hu and Linlin Ge</i>	
GEOUJA System. The First Participation of the University of Jaén at GEOCLEF 2006	913
<i>Manuel García-Vega, Miguel A. García-Cumbreras, L.A. Ureña-López, and José M. Perea-Ortega</i>	
R2D2 at GeoCLEF 2006: A Combined Approach	918
<i>Manuel García-Vega, Miguel A. García-Cumbreras, L. Alfonso Ureña-López, José M. Perea-Ortega, F. Javier Ariza-López, Oscar Ferrández, Antonio Toral, Zornitsa Kozareva, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, Davide Buscaldi, and Paolo Rosso</i>	
MSRA Columbus at GeoCLEF 2006	926
<i>Zhisheng Li, Chong Wang, Xing Xie, Xufa Wang, and Wei-Ying Ma</i>	
Forostar: A System for GIR	930
<i>Simon Overell, João Magalhães, and Stefan Rüger</i>	
NICTA I2D2 Group at GeoCLEF 2006	938
<i>Yi Li, Nicola Stokes, Lawrence Cavedon, and Alistair Moffat</i>	
Blind Relevance Feedback and Named Entity Based Query Expansion for Geographic Retrieval at GeoCLEF 2006	946
<i>Kerstin Bischoff, Thomas Mandl, and Christa Womser-Hacker</i>	
A WordNet-Based Indexing Technique for Geographical Information Retrieval	954
<i>Davide Buscaldi, Paolo Rosso, and Emilio Sanchis</i>	
University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval	958
<i>Claudia Hauff, Dolf Trieschnigg, and Henning Rode</i>	

TALP at GeoCLEF 2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus	962
<i>Daniel Ferrés and Horacio Rodríguez</i>	
GeoCLEF Text Retrieval and Manual Expansion Approaches	970
<i>Ray R. Larson and Fredric C. Gey</i>	
UB at GeoCLEF 2006	978
<i>Miguel E. Ruiz, June Abbas, David Mark, Stuart Shapiro, and Silvia B. Southwick</i>	
The University of Lisbon at GeoCLEF 2006	986
<i>Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade, and Mário J. Silva</i>	
Author Index	995

What Happened in CLEF 2006

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy
carol.peters@isti.cnr.it

Abstract. The organization of the CLEF 2006 evaluation campaign is described and details are provided concerning the tracks, test collections, evaluation infrastructure, and participation.

1 Introduction

The objective of CLEF is to promote research in the field of multilingual system development. This is done through the organisation of annual evaluation campaigns in which a series of tracks designed to test different aspects of mono- and cross-language information retrieval (IR) are offered. The intention is to encourage experimentation with all kinds of multilingual information access – from the development of systems for monolingual retrieval operating on many languages to the implementation of complete multilingual multimedia search services. This has been achieved by offering an increasingly complex and varied set of evaluation tasks over the years. The aim is not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems.

These Proceedings contain descriptions of the experiments conducted within CLEF 2006 – the seventh in a series of annual system evaluation campaigns¹. The main features of the 2006 campaign are briefly outlined in this introductory paper in order to provide the necessary background to the experiments reported in the rest of the Working Notes.

2 Tracks and Tasks in CLEF 2006

CLEF 2006 offered eight tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual textual document retrieval on news collections (Ad Hoc)
- mono- and cross-language information on structured scientific data (Domain-Specific)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval in image collections (ImageCLEF)

¹ CLEF is included in the activities of the DELOS Network of Excellence on Digital Libraries, funded by the Sixth Framework Programme of the European Commission. For information on DELOS, see www.delos.info.

- cross-language speech retrieval (CL-SR)
- multilingual retrieval of Web documents (WebCLEF)
- cross-language geographical retrieval (GeoCLEF)

Although these tracks are the same as those offered in CLEF 2005, many of the tasks offered were new.

Multilingual Text Retrieval (Ad Hoc): Similarly to last year, the 2006 track offered mono- and bilingual tasks on target collections in French, Portuguese, Bulgarian and Hungarian. The topics (i.e. statements of information needs from which queries are derived) were prepared in a wide range of European languages (Bulgarian, English, French, German, Hungarian, Italian, Portuguese, Spanish). We also offered a bilingual task aimed at encouraging system testing with non-European languages against an English target collection. Topics were supplied in: Amharic, Chinese, Hindi, Indonesian, Oromo and Telugu. This choice of languages was determined by the demand from participants. In addition, a new robust task was offered; this task emphasized the importance of stable performance over languages instead of high average performance in mono-, bilingual and multilingual IR. It made use of test collections previously developed at CLEF. The track was coordinated jointly by ISTI-CNR and U.Padua (Italy) and U.Hildesheim (Germany).

Cross-Language Scientific Data Retrieval (Domain-Specific): This track studied retrieval in a domain-specific context using the GIRT-4 German/English social science database and two Russian corpora: Russian Social Science Corpus (RSSC) and the ISSS collection for sociology and economics. Multilingual controlled vocabularies (German-English, English-German, German-Russian, English-Russian) were available. Monolingual and cross-language tasks were offered. Topics were prepared in English, German and Russian. Participants could make use of the indexing terms inside the documents and/or the social science thesaurus provided, not only as translation means but also for tuning relevance decisions of their system. The track was coordinated by IZ Bonn (Germany).

Interactive CLIR (iCLEF): For CLEF 2006, the interactive track joined forces with the image track to work on a new type of interactive image retrieval task to better capture the interplay between image and the multilingual reality of the internet for the public at large. The task was based on the popular image perusal community Flickr (www.flickr.com), a dynamic and rapidly changing database of images with textual comments, captions, and titles in many languages and annotated by image creators and viewers cooperatively in a self-organizing ontology of tags (a so-called “folksonomy”). The track was coordinated by UNED (Spain), U. Sheffield (UK) and SICS (Sweden).

Multilingual Question Answering (QA@CLEF): This track, which has received increasing interest at CLEF since 2003, evaluated both monolingual (non-English) and cross-language QA systems. The main task evaluated open domain QA systems. Target collections were offered in Bulgarian, Dutch, English (bilingual only), French, German, Italian, Portuguese and Spanish. In addition, three pilot tasks were organized: a task that assessed question answering using Wikipedia, the online encyclopaedia; an Answer Validation exercise; and a “Time-constrained” exercise that was conducted

in the morning previous to the workshop. A number of institutions (one for each language) collaborated in the organization of the main task; the Wikipedia activity was coordinated by U. Amsterdam (The Netherlands), the Answer Validation exercise by UNED (Spain) and the Time-constrained exercise by U. Alicante (Spain). The overall coordination of this track was by ITC-irst and CELCT, Trento (Italy).

Cross-Language Retrieval in Image Collections (ImageCLEF): This track evaluated retrieval of images described by text captions based on queries in a different language; both text and image matching techniques were potentially exploitable. Two main sub-tracks were organised for photographic and medical image retrieval. Each track offered two tasks: bilingual ad hoc retrieval (collection in English, queries in a range of languages) and an annotation task in the first case; medical image retrieval (collection with casenotes in English, French and German, queries derived from short text plus image - visual, mixed and semantic queries) and automatic annotation for medical images (fully categorized collection, categories available in English and German) in the second. The tasks offered different and challenging retrieval problems for cross-language image retrieval. Image analysis was not required for all tasks and a default visual image retrieval system was made available for participants as well as results from a basic text retrieval system. The track coordinators were University of Sheffield (UK) and the University and Hospitals of Geneva (Switzerland). Oregon Health and Science University (USA), Victoria University, Melbourne (Australia), RWTH Aachen University (Germany), and Vienna University of Technology (Austria) collaborated in the task organization.

Cross-Language Speech Retrieval (CL-SR): In 2005, the CL-SR track built a reusable test collection for searching spontaneous conversational English speech using queries in five languages (Czech, English, French, German and Spanish), speech recognition for spoken words, manually and automatically assigned controlled vocabulary descriptors for concepts, dates and locations, manually assigned person names, and hand-written segment summaries. The 2006 CL-SR track included a second test collection containing about 500 hours of Czech speech. Multilingual topic sets were again created for five languages. The track was coordinated by the University of Maryland (USA) and Dublin City University (Ireland).

Multilingual Web Retrieval (WebCLEF): WebCLEF 2006 used the EuroGOV collection, with web pages crawled from European governmental sites for over 20 languages/countries. It was decided to focus this year on the mixed-monolingual known-item topics. The topics were a mixture of old topics and new topics. The old topics were a subset of last year's topics; the new topics were provided by the organizers, using a new method for generating known-item test beds and some human generated new topics. The experiments explored two complementary dimensions: old vs new topics; topics generated by participants vs automatically topics generated by the organizers

Cross-Language Geographical Retrieval (GeoCLEF): The track provided a framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects. Participants were offered a TREC style ad hoc retrieval task based on existing CLEF collections. The aim was to compare methods of query translation, query expansion, translation of geographical references, use of text and

spatial retrieval methods separately or combined, retrieval models and indexing methods. Given a multilingual statement describing a spatial user need (topic), the challenge was to find relevant documents from target collections in English, Portuguese, German or Spanish. Monolingual and cross-language tasks were activated. Topics were offered in English, German, Portuguese, Spanish and Japanese. Spatial analysis was not required to participate in this task but could be used to augment text-retrieval methods. A number of groups collaborated in the organization of the track; the overall coordination was by UC Berkley (USA) and U. Sheffield (UK).

3 Test Collections

The CLEF test collections, created as a result of the evaluation campaigns, consist of topics or queries, documents, and relevance assessments. Each track was responsible for preparing its own topic/query statements and for performing relevance assessments of the results submitted by participating groups. A number of different document collections were used in CLEF 2006 to build the test collections:

- CLEF multilingual comparable corpus of more than 2 million news docs in 12 languages (see Table 1) ; this corpus was unchanged from 2005. Parts of this collection were used in three tracks: Ad-Hoc (all languages except Finnish, Swedish and Russian), Question Answering (all languages except Finnish, Hungarian, Swedish and Russian) and GeoCLEF (English, German, Portuguese and Spanish).
- The CLEF domain-specific collection consisting of the GIRT-4 social science database in English and German (over 300,000 documents) and two Russian databases: the Russian Social Science Corpus (approx. 95,000 documents) and the Russian ISISS collection for sociology and economics (approx. 150,000 docs). The ISISS corpus was new this year. Controlled vocabularies in German-English and German-Russian were also made available to the participants in this track. This collection was used in the domain-specific track.
- The ImageCLEF track used four collections:
 - the ImageCLEFmed radiological medical database based on a dataset containing images from the Casimage, MIR, PEIR, and PathoPIC datasets (about 50,000 images) with case notes in English (majority) but also German and French.
 - the IRMA collection in English and German of 10,000 images for automatic medical image annotation
 - the IAPR TC-12 database of 25,000 photographs with captions in English, German and Spanish
 - a general photographic collection for image annotation provided by LookThatUp (LTUtech) database
- The Speech retrieval track used the Malach collection of spontaneous conversational speech derived from the Shoah archives in English (more than 750 hours) and Czech (approx 500 hours).

Table 1. Sources and dimensions of the CLEF 2006 multilingual comparable corpus

Collection	Added in	Size (MB)	No. of Docs	Median Size of Docs. (Tokens)
Bulgarian: Sega 2002	2005	120	33,356	NA
Bulgarian: Standart 2002	2005	93	35,839	NA
Dutch: Algemeen Dagblad 94/95	2001	241	106483	166
Dutch: NRC Handelsblad 94/95	2001	299	84121	354
English: LA Times 94	2000	425	113005	421
English: Glasgow Herald 95	2003	154	56472	343
Finnish: Aamulehti late 94/95	2002	137	55344	217
French: Le Monde 94	2000	158	44013	361
French: ATS 94	2001	86	43178	227
French: ATS 95	2003	88	42615	234
German: Frankfurter Rundschau94	2000	320	139715	225
German: Der Spiegel 94/95	2000	63	13979	213
German: SDA 94	2001	144	71677	186
German: SDA 95	2003	144	69438	188
Hungarian: Magyar Hirlap 2002	2005	105	49,530	NA
Italian: La Stampa 94	2000	193	58051	435
Italian: AGZ 94	2001	86	50527	187
Italian: AGZ 95	2003	85	48980	192
Portuguese: Público 1994	2004	164	51751	NA
Portuguese: Público 1995	2004	176	55070	NA
Portuguese: Folha 94	2005	108	51,875	NA
Portuguese: Folha 95	2005	116	52,038	NA
Russian: Izvestia 95	2003	68	16761	NA
Spanish: EFE 94	2001	511	215738	290
Spanish: EFE 95	2003	577	238307	299
Swedish: TT 94/95	2002	352	142819	183

SDA/ATS/AGZ = Schweizerische Depeschagentur (Swiss News Agency)

EFE = Agencia EFE S.A (Spanish News Agency)

TT = Tidningarnas Telegrambyrå (Swedish newspaper)

- The WebCLEF track used a collection crawled from European governmental sites, called EuroGOV. This collection consists of more than 3.35 million pages from 27 primary domains. The most frequent languages are Finnish (20%), German (18%), Hungarian (13%), English (10%), and Latvian (9%).

4 Technical Infrastructure

The CLEF technical infrastructure is managed by the DIRECT system. DIRECT manages the test data plus results submission and analyses for the ad hoc, question answering and geographic IR tracks. It has been designed to facilitate data management tasks but also to support the production, maintenance, enrichment and interpretation of the scientific data for subsequent in-depth evaluation studies.

The technical infrastructure is thus responsible for:

- the track set-up, harvesting of documents, management of the registration of participants to tracks;
- the submission of experiments, collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- the provision of common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;
- the provision of common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses.

DIRECT was designed and implemented by Giorgio Di Nunzio and Nicola Ferro and is described in more detail in a paper in these Proceedings.

5 Participation

A total of 90 groups submitted runs in CLEF 2006, as opposed to the 74 groups of CLEF 2005: 59.5(43) from Europe, 14.5(19) from N.America; 10(10) from Asia, 4(1) from S.America and 2(1) from Australia². Last year's figures are given between brackets. The breakdown of participation of groups per track is as follows: Ad Hoc 25; Domain-Specific 4; iCLEF 3; QAatCLEF 37; ImageCLEF 25; CL-SR 6; WebCLEF 8; GeoCLEF 17. As in previous years, participating groups consisted of a nice mix of new-comers (34) and groups that had participated in one or more previous editions (56). Most of the groups came from academia; there were just 9 research groups from industry. A list of groups and indications of the tracks in which they participated is given in Appendix to the online Working Notes which can be found at www.clef-campaign.org under Publications.

Figure 1 shows the growth in participation over the years and Figure 2 shows the shift in focus over the years as new tracks have been added.

² The 0.5 figures result from a Mexican/Spanish collaboration.

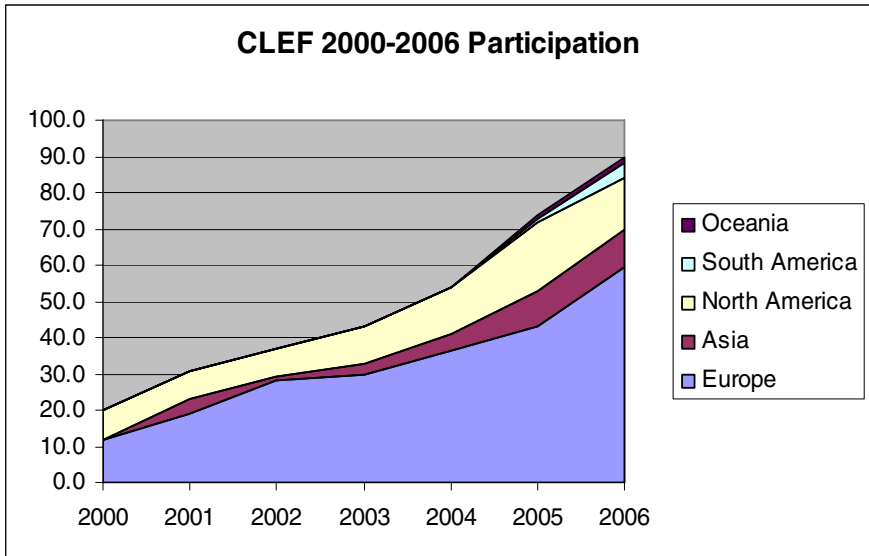


Fig. 1. CLEF 2000 – 2006: Increase in Participation

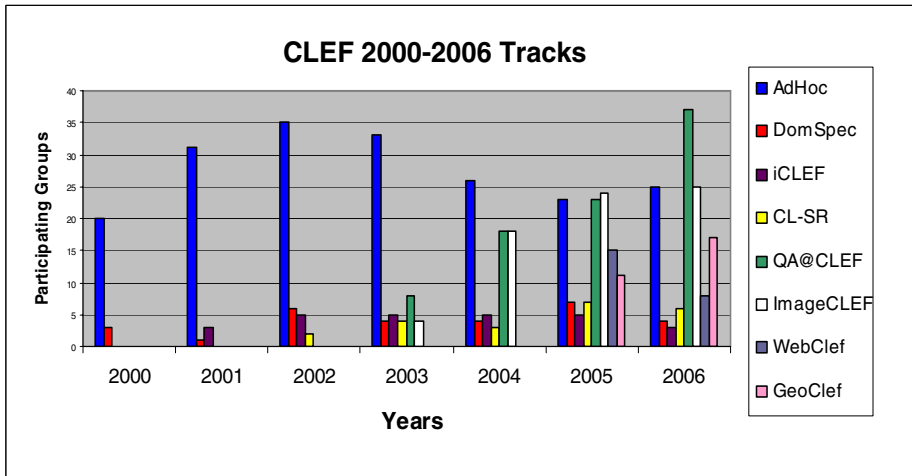


Fig. 2. CLEF 2000 – 2006: Shift in Participation

6 Main Results

The results achieved by CLEF in these seven years of activity are impressive. We can summarise them in the following main points:

- documented improvement in system performance for cross-language text retrieval systems
- R&D activity in new areas such as cross-language question answering, multilingual retrieval for mixed media, and cross-language geographic information retrieval
- an extensive body of research literature on multilingual information access issues
- creation of important, reusable test collections for system benchmarking
- building of a strong, multidisciplinary research community.

Furthermore, CLEF evaluations have provided qualitative and quantitative evidence along the years as to which methods give the best results in certain key areas, such as multilingual indexing, query translation, resolution of translation ambiguity, results merging.

However, although CLEF has done much to promote the development of multilingual IR systems, so far the focus has been on building and testing research prototypes rather than developing fully operational systems. There is still a considerable gap between the research and the application communities and, despite the strong demand for and interest in multilingual IR functionality, there are still very few commercially viable systems on offer. The challenge that CLEF must face in the near future is how to best transfer the research results to the market place. CLEF 2006 took a first step in this direction with the organization of the real time exercise as part of the question-answering track and this issue was discussed in some detail at the workshop.

In our opinion, if the gap between academic excellence and commercial adoption of CLIR technology is to be bridged, we need to extend the current CLEF formula in order to give application communities the possibility to benefit from the CLEF evaluation infrastructure without the need to participate in academic exercises that may be irrelevant to their current needs. We feel that CLEF should introduce an application support structure aimed at encouraging take-up of the technologies tested and optimized within the context of the evaluation exercises. This structure would provide tools, resources, guidelines and consulting services to applications or industries that need to include multilingual functionality within a service or product. We are currently considering ways in which to implement an infrastructure of this type.

Acknowledgements

I could not run the CLEF evaluation initiative and organize the annual workshops without considerable assistance from many people. CLEF is organized on a distributed basis, with different research groups being responsible for the running of the various tracks. My gratitude goes to all those who were involved in the coordination of the 2006 campaign. It is really impossible for me to list the names of all the people who helped in the organization of the different tracks. Here below, let me just mention those responsible for the overall coordination:

- Giorgio Di Nunzio, Nicola Ferro and Thomas Mandl for the Ad Hoc Track
- Maximilian Stempfhuber, Stefan Baerisch and Natalia Loukachevitch for the Domain-Specific track
- Julio Gonzalo, Paul Clough and Jussi Karlgren for iCLEF

- Bernardo Magnini, Danilo Giampiccolo, Fernando Llopis, Elisa Noguera, Anselmo Peñas and Maarten de Rijke for QA@CLEF
- Paul Clough, Thomas Deselaers, Michael Grubinger, Allan Hanbury, William Hersch, Thomas Lehmann and Henning Müller for ImageCLEF
- Douglas W. Oard and Gareth J. F. Jones for CL-SR
- Krisztian Balog, Leif Azzopardi, Jaap Kamps and Maarten de Rijke for Web-CLEF
- Fredric Gey, Ray Larson and Mark Sanderson as the main coordinators of GeoCLEF

I apologise for those I have not mentioned here. However, I really must express my appreciation to Diana Santos and her colleagues at Linguatca in Norway and Portugal, for all their efforts aimed at supporting the inclusion of Portuguese in CLEF activities. I also thank all those colleagues who have helped us by preparing topic sets in different languages and in particular the NLP Lab., Dept. of Computer Science and Information Engineering of the National Taiwan University and the Institute for Information Technology, Hyderabad, India, for their work on Chinese and Hindi, respectively. I should also like to thank the members of the CLEF Steering Committee who have assisted me with their advice and suggestions throughout this campaign.

Furthermore, I gratefully acknowledge the support of all the data providers and copyright holders, and in particular:

- The Los Angeles Times, for the American-English data collection
- SMG Newspapers (The Herald) for the British-English data collection
- Le Monde S.A. and ELDA: Evaluations and Language resources Distribution Agency, for the French data
- Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections
- InformationsZentrum Sozialwissen-schaften, Bonn, for the GIRT database
- SocioNet system for the Russian Social Science Corpora
- Hypersystems Srl, Torino and La Stampa, for the Italian data
- Agencia EFE S.A. for the Spanish data
- NRC Handelsblad, Algemeen Dagblad and PCM Landelijke dagbladen/Het Parool for the Dutch newspaper data
- Aamulehti Oyj and Sanoma Osakeyhtiö for the Finnish newspaper data
- Russika-Izvestia for the Russian newspaper data
- Público, Portugal, and Linguatca for the Portuguese (PT) newspaper collection
- Folha, Brazil, and Linguatca for the Portuguese (BR) newspaper collection
- Tidningarnas Telegrambyrå (TT) SE-105 12 Stockholm, Sweden for the Swedish newspaper data
- Schweizerische Depeschagentur, Switzerland, for the French, German and Italian Swiss news agency data
- Ringier Kiadoi Rt. [Ringier Publishing Inc.] and the Research Institute for Linguistics, Hungarian Acad. Sci. for the Hungarian newspaper documents
- Sega AD, Sofia; Standart Nyuz AD, Sofia, and the BulTreeBank Project, Linguistic Modelling Laboratory, IPP, Bulgarian Acad. Sci, for the Bulgarian newspaper documents

- St Andrews University Library for the historic photographic archive
- University and University Hospitals, Geneva, Switzerland and Oregon Health and Science University for the ImageCLEFmed Radiological Medical Database
- Aachen University of Technology (RWTH), Germany for the IRMA database of annotated medical images
- The Survivors of the Shoah Visual History Foundation, and IBM for the Malach spoken document collection

Without their contribution, this evaluation activity would be impossible.

Last and not least, I should like to express my gratitude to Alessandro Nardi in Pisa and José Luis Vicedo, Patricio Martínez Barco and Maximiliano Saiz Noeda, U. Alicante, for their assistance in the organisation of the CLEF 2006 Workshop.

Scientific Data of an Evaluation Campaign: Do We Properly Deal with Them?

Maristella Agosti, Giorgio Maria Di Nunzio, and Nicola Ferro

Department of Information Engineering – University of Padua
Via Gradenigo, 6/b – 35131 Padova – Italy
{agosti, dinunzio, ferro}@dei.unipd.it

Abstract. This paper examines the current way of keeping the data produced during the evaluation campaigns and highlights some shortenings of it. As a consequence, we propose a new approach for improving the management evaluation campaigns’ data. In this approach, the data are considered as scientific data to be cured and enriched in order to give full support to longitudinal statistical studies and long-term preservation.

1 Introduction

When we reason about the data and information produced during an evaluation campaign, we should be aware that a lot of valuable *scientific data* are produced [1,2]. Indeed, if we consider some of the outcomes of an evaluation campaign, we can see how they actually are different kinds of scientific data:

- *experiments*: the primary scientific data produced by the participants of an evaluation campaign represent the starting point for any subsequent analysis;
- *performance measurements*: metrics, such as precision and recall, are derived from the experiments and are used to evaluate the performances of an *Information Retrieval System (IRS)*
- *descriptive statistics*: the statistics, such as mean or median, are computed from the performance measurements and summarized the overall performances of an IRS or a group of IRSs;
- *statistical analyses*: different statistical techniques, such as hypothesis test, makes use of the performance measurements and the descriptive statistics in order to perform an in-depth analysis of the experiments and assess their differences.

A huge amount of the above mentioned data is produced each year during an evaluation campaign and these data are an integral part of the scientific research in the information retrieval field.

When we deal with scientific data, “the *lineage (provenance)* of the data must be tracked, since a scientist needs to know where the data came from [...] and what cleaning, rescaling, or modelling was done to arrive at the data to be interpreted” [3]. In addition, [4] points out how provenance is “important

in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information”.

Furthermore, when scientific data are maintained for further and future use, they should be enriched and, besides information about provenance, also changes at sources occurred over time need to be tracked. Sometimes the enrichment of a portion of scientific data can make use of a *citation* for explicitly mentioning and making references to useful information.

In this paper we examine whether the current methodology properly deals with the data produced during an evaluation campaign by recognizing that they are in effect valuable scientific data. Furthermore, we describe the *data curation approach* [5,6] which we have undertaken to overcome some of the shortenings of the current methodology and we have applied in designing and developing the infrastructure for the *Cross-Language Evaluation Forum (CLEF)*.

The paper is organized as follows: Section 2 introduces the motivations and the objectives of our research work; Section 3 describes the work carried out in developing the CLEF infrastructure; finally, Section 4 draws some conclusions.

2 Motivations and Objectives

2.1 Experimental Collections

Nowadays, the experimental evaluation is carried out in important international evaluation forums, such as *Text REtrieval Conference (TREC)*, CLEF, and *NII-NACSIS Test Collection for IR Systems (NTCIR)*, which bring research groups together, provide them with the means for measuring the performances of their systems, discuss and compare their work.

All of the previously mentioned initiatives are generally carried out according to the Cranfield methodology, which makes use of *experimental collections* [7]. An experimental collection is a triple $\mathcal{C} = (D, Q, J)$, where: D is a set of documents, called also collection of documents; Q is a set of topics, from which the actual queries are derived; J is a set relevance judgements, i.e. for each topic $q \in Q$ the documents $d \in D$, which are relevant for the topic q , are determined. An experimental collection $\mathcal{C}$ allows the comparison of two retrieval methods, say X and Y , according to some measurements which quantifies the retrieval performances of these methods. An experimental collection both provides a common test-bed to be indexed and searched by the IRS X and Y and guarantees the possibility of replicating the experiments.

Nevertheless, the Cranfield methodology is mainly focused on creating comparable experiments and evaluating the performances of an IRS rather than modeling and managing the scientific data produced during an evaluation campaign.

As an example, note that the exchange of information between organizers and participants is mainly performed by means of textual files formatted according to the TREC data format, which is the de-facto standard in this field. Note that this information represents a first kind of scientific data produced during the

evaluation process. The following is a fragment of the results of an experiment submitted by a participant to the organizers, where the gray header is not really present in the exchanged data but serves here as an explanation of the fields.

Topic	Iter.	Document	Rank	Score	Experiment
141	Q0	AGZ.950609.0067	0	0.440873414278	IMSMIPO
141	Q0	AGZ.950613.0165	1	0.305291658641	IMSMIPO
...					

In the above data, each row represents a record of an experiment, where fields are separated by white spaces. There is the field which specifies the unique identifier of the topic (e.g. 141), the field for the unique identifier of the document (e.g. AGZ.950609.0067), the field which identifies the experiment (e.g. IMSMFPO), and so on, as specified by the gray headers.

As it can be noted from the above examples, this format is suitable for a simple data exchange between participants and organizers. Nevertheless, neither this format provides any metadata explaining its content nor a scheme exists in order to define the structure of each file, the data type of each field, and various constraints on the data, such as numeric floating point precision. In addition, this format does not ensure that any kind of constraint is complied with, e.g. we would avoid to retrieve the same document twice or more for the same topic. Finally, this format is not very suitable for modelling the information space involved by an evaluation forum because the relationships among the different entities (documents, topics, experiments, participants) are not modeled and each entity is treated separately from the others.

Furthermore, present collections keeping over time does not permit systematic studies on reached improvements by participants over the years, for example in a specific multilingual setting [8].

We argue that the information space implied by an evaluation forum needs an appropriate conceptual model which takes into consideration and describes all the entities involved by the evaluation forum. In fact, an appropriate conceptual model is the necessary basis to make the scientific data produced during the evaluation an active part of all those information enrichments, as data provenance and citation, we have described in the previous section. This conceptual model can be also translated into an appropriate logical model in order to manage the information of an evaluation forum by using the database technology, as an example. Finally, from this conceptual model we can derive also appropriate data formats for exchanging information among organizers and participants, such as an *eXtensible Markup Language (XML)* format that complies with an XML Schema [9,10] which describes and constraints the exchanged information.

2.2 Statistical Analysis of Experiments

The Cranfield methodology is mainly focused on how to evaluate the performances of two systems and how to provide a common ground which makes the experimental results comparable. [11] points out that, in order to evaluate retrieval

performances, we do not need only an experimental collection and measures for quantifying retrieval performances, but also a statistical methodology for judging whether measured differences between retrieval methods X and Y can be considered statistically significant.

To address this issue, evaluation forums have traditionally carried out statistical analyses, which provide participants with an overview analysis of the submitted experiments, as in the case of the overview papers of the different tracks at TREC and CLEF; some recent examples of this kind of papers are [12,13]. Furthermore, participants may conduct statistical analyses on their own experiments by using either ad-hoc packages, such as IR-STAT-PAK¹, or generally available software tools with statistical analysis capabilities, like R², SPSS³, or MATLAB⁴. However, the choice of whether performing a statical analysis or not is left up to each participant who may even not have all the skills and resources needed to perform such analyses. Moreover, when participants perform statistical analyses using their own tools, the comparability among these analyses could not be fully granted because, for example, different statistical tests can be employed to analyze the data, or different choices and approximations for the various parameters of the same statistical test can be made.

Thus, we can observe that, in general, there is a limited support to the systematical employment of statistical analysis by participants. For this reason, we suggest that evaluation forums should support and guide participants in adopting a more uniform way of performing statistical analyses on their own experiments. In this way, participants can not only benefit from standard experimental collections which make their experiments comparable, but they can also exploit standard tools for the analysis of the experimental results, which make the analysis and assessment of their experiments comparable too.

2.3 Information Enrichment and Interpretation

As introduced in Section 1, scientific data, their enrichment and interpretation are essential components of scientific research. The Cranfield methodology traces out how these scientific data have to be produced, while the statistical analysis of experiments provide the means for further elaborating and interpreting the experimental results. Nevertheless, the current methodologies does not require any particular coordination or synchronization between the basic scientific data and the analyses on them, which are treated as almost separated items. On the contrary, researchers would greatly benefit from an integrated vision of them, where the access to a scientific data item could also offer the possibility of retrieving all the analyses and interpretations on it. Furthermore, it should be possible to enrich the basic scientific data in an incremental way, progressively adding further analyses and interpretations on them.

¹ <http://users.cs.dal.ca/~jamie/pubs/IRSP-overview.html>

² <http://www.r-project.org/>

³ <http://www.spss.com/>

⁴ <http://www.mathworks.com/>

Let us consider what is currently done in an evaluation forum:

- Experimental collections:
 - there are few or no metadata about document collections, the context they refer to, how they have been created, and so on;
 - there are few or no metadata about topics, how they have been created, the problems encountered by their creators, what documents creators found relevant for a given topic, and so on;
 - there are few or no metadata about how pools have been created and about the relevance assessments, the problems which have been faced by the assessors when dealing with difficult topics;
- Experiments:
 - there are few or no metadata about them, such as what techniques have been adopted or what tunings have been carried out;
 - they can be not publicly accessible, making it difficult for other researchers to make a direct comparison with their own experiments;
 - their citation can be an issue;
- Performance measurements:
 - there are no metadata about how a measure has been created, which software has been used to compute it, and so on;
 - often only summaries are publicly available while all the detailed measurements may be not accessible;
 - their format can be not suitable for further computer processing;
 - their modelling and management needs to be dealt with;
- Descriptive statistics and hypothesis tests:
 - they are mainly limited to task overviews produced by organizers;
 - participants may not have all the skills needed to perform a statistical analysis;
 - participants can carry out statistical analyses only on their own experiments without the possibility of comparing them with the experiments of other participants;
 - the comparability among the statistical analyses conducted by the participants is not fully granted due to possible differences in the design of the statistical experiments.

These issues are better faced and framed in the wider context of the *curation of scientific data*, which plays an important role on the systematic definition of a proper methodology to manage and promote the use of data. The e-Science Data Curation Report gives the following definition of data curation [14]: “the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose”.

This definition implies that we have to take into consideration the possibility of information enrichment of scientific data, meant as archiving and preserving scientific data so that the experiments, records, and observations will be available for future research, as well as provenance, curation, and citation of scientific data

items. The benefits of this approach include the growing involvement of scientists in international research projects and forums and increased interest in comparative research activities. Furthermore, the definition introduced above reflects the importance of some of the many possible reasons for which keeping data is important, for example: re-use of data for new research, including collection based research to generate new science; retention of unique observational data which is impossible to re-create; retention of expensively generated data which is cheaper to maintain than to re-generate; enhancing existing data available for research projects; validating published research results.

As a concrete example in the field of information retrieval, please consider the data fusion problem [15], where lists of results produced by different systems have to be merged into a single list. In this context, researchers do not start from scratch, but they often experiment their merging algorithms by using the list of results produced in experiments carried out even by other researchers. This is the case, for example, of the CLEF 2005 multilingual merging track [12], which provided participants with some of the CLEF 2003 multilingual experiments as list of results to be used as input to their merging algorithms. It is now clear that researchers of this field would benefit by a clear data curation strategy, which promotes the re-use of existing data and allows the data fusion experiments to be traced back to the originary list of results and, perhaps, to the analyses and interpretations about them.

Thus, we consider all these points as requirements that should be taken into account when we are going to produce and manage scientific data that come out from the experimental evaluation of an IRS. In addition, to achieve the full information enrichment discussed in Section 1, both the experimental datasets and their further elaboration, such as their statistical analysis, should be first class objects that can be directly referenced and cited. Indeed, as recognized by [14], the possibility of citing scientific data and their further elaboration is an effective way for making scientists and researchers an active part of the digital curation process.

3 The CLEF Infrastructure

3.1 Conceptual Model for an Evaluation Forum

As discussed in the previous section, we need to design and develop a proper conceptual model of the information space involved by an evaluation forum. Indeed, this conceptual model provide us with the basis needed to offer all the information enrichment and interpretation features described above.

Figure 1 shows the *Entity-Relationship (ER)* schema which represents the conceptual model we have developed. The conceptual model is built around five main areas of modelling:

- **evaluation forum**: deals with the different aspects of an evaluation forum, such as the conducted evaluation campaigns and the different editions of each campaign, the tracks along which the campaign is organized, the subscription of the participants to the tracks, the topics of each track;

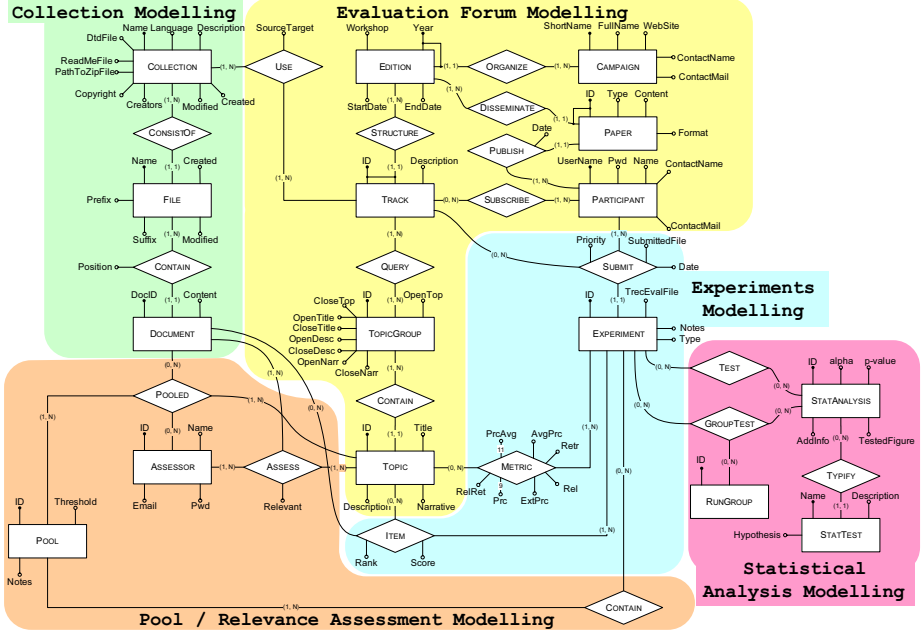


Fig. 1. Conceptual model for the information space of an evaluation forum

- **collection**: concerns the different collections made available by an evaluation forum; each collection can be organized into various files and each file may contain one or more multimedia documents; the same collection can be used by different tracks and by different editions of the evaluation campaign;
- **experiments**: regards the experiments submitted by the participants and the evaluation metrics computed on those experiments, such as precision and recall;
- **pool/relevance assessment**: is about the pooling method [16], where a set of experiments is pooled and the documents retrieved in those experiments are assessed with respect to the topics of the track the experiments belongs to;
- **statistical analysis**: models the different aspects concerning the statistical analysis of the experimental results, such as the type of statistical test employed, its parameters, the observed test statistic, and so forth.

3.2 DIRECT: The Running Prototype

The proposed approach has been implemented in a prototype, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [17,18,19], and it has been tested in the context of the CLEF 2005 and 2006 evaluation campaigns. The initial prototype moves a first step towards a data curation approach to evaluation initiatives, by providing support for:

- the management of an evaluation forum: the track set-up, the harvesting of documents, the management of the subscription of participants to tracks;
- the management of submission of experiments, the collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;
- common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses;
- common XML format for exchanging data between organizers and participants.

DIRECT was successfully adopted during the CLEF 2005 campaign. It was used by nearly 30 participants spread over 15 different nations, who submitted more than 530 experiments; then 15 assessors assessed more than 160,000 documents in seven different languages, including Russian and Bulgarian which do not have a latin alphabet. During the CLEF 2006 campaign, it has been used by nearly 75 participants spread over 25 different nations, who have submitted around 570 experiments; 40 assessors assessed more than 198,500 documents in nine different languages. DIRECT was then used for producing reports and overview graphs about the submitted experiments [20,21].

DIRECT has been developed by using the Java⁵ programming language, which ensures great portability of the system across different platforms. We used the PostgreSQL⁶ *DataBase Management System (DBMS)* for performing the actual storage of the data. Finally, all kinds of *User Interface (UI)* in DIRECT are Web-based interfaces, which make the service easily accessible to end-users without the need of installing any kind of software. These interfaces have been developed by using the Apache STRUTS⁷ framework, an open-source framework for developing Web applications.

4 Conclusions

The discussed data curation approach that can help to face the test-collection challenge for the evaluation and future development of information access and extraction components of interactive information management systems. On the basis of the experience gained keeping and managing the data of interest of the CLEF evaluation campaign, we are testing the considered requirements to revise the proposed approach.

A prototype of the carrying out the proposed approach, called DIRECT, has been implemented and widely tested during the CLEF 2005 and 2006 evaluation campaigns. On the basis of the experience gained, we are enhancing the proposed conceptual model and architecture, in order to implement a second version of the prototype to be tested and validated during the next CLEF campaigns.

⁵ <http://java.sun.com/>

⁶ <http://www.postgresql.org/>

⁷ <http://struts.apache.org/>

Acknowledgements

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

1. Agosti, M., Di Nunzio, G.M., Ferro, N.: The Importance of Scientific Data Curation for Evaluation Campaigns. In: DELOS Conference 2007 Working Notes, ISTI-CNR, Gruppo ALI, Pisa, Italy, pp. 185–193 (2007)
2. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In: Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), National Institute of Informatics, Tokyo, Japan, pp. 62–73 (2007)
3. Abiteboul, S., Agrawal, R., Bernstein, P., Carey, M., Ceri, S., Croft, B., DeWitt, D., Franklin, M., Garcia-Molina, H., Gawlick, D., Gray, J., Haas, L., Halevy, A., Hellerstein, J., Ioannidis, Y., Kersten, M., Pazzani, M., Lesk, M., Maier, D., Naughton, J., Schek, H.J., Sellis, T., Silberschatz, A., Stonebraker, M., Snodgrass, R., Ullman, J.D., Weikum, G., Widom, J., Zdonik, S.: The Lowell Database Research Self-Assessment. *Communications of the ACM (CACM)* 48, 111–118 (2005)
4. Ioannidis, Y., Maier, D., Abiteboul, S., Buneman, P., Davidson, S., Fox, E.A., Halevy, A., Knoblock, C., Rabitti, F., Schek, H.J., Weikum, G.: Digital library information-technology infrastructures. *International Journal on Digital Libraries* 5, 266–274 (2005)
5. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Data Curation Approach to Support In-depth Evaluation Studies. In: Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006), pp. 65–68 (2006) (last visited, March 23, 2007), <http://ucdata.berkeley.edu/sigir2006-mlia.htm>
6. Agosti, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF: Ongoing Activities and Plans for the Future. In: Proc. 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, National Institute of Informatics, Tokyo, Japan, pp. 493–504 (2007)
7. Cleverdon, C.W.: The Cranfield Tests on Index Languages Devices. In: *Readings in Information Retrieval*, pp. 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997)
8. Agosti, M., Di Nunzio, G.M., Ferro, N.: Evaluation of a Digital Library System. In: Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, pp. 73–78 (2004) (last visited, March 23, 2007), http://dlib.ionio.gr/wp7/workshop2004_program.html
9. W3C: XML Schema Part 1: Structures - W3C Recommendation 28 October 2004. (2004) (last visited, March 23, 2007), <http://www.w3.org/TR/xmlschema-1/>
10. W3C: XML Schema Part 2: Datatypes - W3C Recommendation 28 October 2004. (2004) (last visited, March 23, 2007), <http://www.w3.org/TR/xmlschema-2/>
11. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), pp. 329–338. ACM Press, New York, USA (1993)

12. Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: Ad Hoc Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 11–36. Springer, Heidelberg (2006)
13. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005), National Institute of Standards and Technology (NIST), Special Publication 500-266, Washington, USA (2005) (last visited, March 23, 2007), http://trec.nist.gov/pubs/trec14/t14_proceedings.html
14. Lord, P., Macdonald, A.: e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. The JISC Committee for the Support of Research (JCSR) (2003) (last visited, March 23, 2007), http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
15. Croft, W.B.: Combining Approaches to Information Retrieval. In: Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, pp. 1–36. Kluwer Academic Publishers, Norwell (MA), USA (2000)
16. Harman, D.K.: Overview of the First Text REtrieval Conference (TREC-1). In The First Text REtrieval Conference (TREC-1), National Institute of Standards and Technology (NIST), Special Publication 500-207, Washington, USA (1992) (last visited, March 23, 2007), <http://trec.nist.gov/pubs/trec1/papers/01.txt>
17. Di Nunzio, G.M., Ferro, N.: DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns. In: Proc. 8th DELOS Thematic Workshop on Future Digital Library Management Systems: System Architecture and Information Access, pp. 58–63 (2005) (last visited, March 23, 2007), <http://dbis.cs.unibas.ch/delos-website/delos-dagstuhl-handout-all.pdf>
18. Di Nunzio, G.M., Ferro, N.: DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 483–484. Springer, Heidelberg (2005)
19. Di Nunzio, G.M., Ferro, N.: Scientific Evaluation of a DLMS: a service for evaluating information access components. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 536–539. Springer, Heidelberg (2006)
20. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks and Domain-Specific Tracks. In: Peters, C., Quochi, V. (eds.) Working Notes for the CLEF 2005 Workshop (2005) (last visited, March 23, 2007), http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appen_dix_a.pdf
21. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the Ad-hoc Bilingual and Monolingual Tasks. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Working Notes for the CLEF 2006 Workshop (2006) (last visited, March 23, 2007), http://www.clef-campaign.org/2006/working_notes/workingnotes2006/Appendix_Ad-Hoc.pdf

CLEF 2006: Ad Hoc Track Overview

Giorgio M. Di Nunzio¹, Nicola Ferro¹, Thomas Mandl², and Carol Peters³

¹ Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it

² Information Science, University of Hildesheim – Germany
mandl@uni-hildesheim.de

³ ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2006 ad hoc track and discuss the main characteristics of the tasks offered to test monolingual, bilingual, and multilingual textual document retrieval systems. The track was divided into two streams. The main stream offered mono- and bilingual tasks using the same collections as CLEF 2005: Bulgarian, English, French, Hungarian and Portuguese. The second stream, designed for more experienced participants, offered the so-called "robust task" which used test collections from previous years in six languages (Dutch, English, French, German, Italian and Spanish) with the objective of privileging experiments which achieve good stable performance over all queries rather than high average performance. The performance achieved for each task is presented and the results are commented. The document collections used were taken from the CLEF multilingual comparable corpus of news documents.

1 Introduction

The ad hoc retrieval track is generally considered to be the core track in the *Cross-Language Evaluation Forum (CLEF)*. The aim of this track is to promote the development of monolingual and cross-language textual document retrieval systems. The CLEF 2006 ad hoc track was structured in two streams. The main stream offered monolingual tasks (querying and finding documents in one language) and bilingual tasks (querying in one language and finding documents in another language) using the same collections as CLEF 2005. The second stream, designed for more experienced participants, was the "robust task", aimed at finding relevant documents for difficult queries. It used test collections developed in previous years.

The **Monolingual** and **Bilingual** tasks were principally offered for Bulgarian, French, Hungarian and Portuguese target collections. Additionally, in the bilingual task only, newcomers (i.e. groups that had not previously participated in a CLEF cross-language task) or groups using a "new-to-CLEF" query language could choose to search the English document collection. The aim in all

cases was to retrieve relevant documents from the chosen target collection and submit the results in a ranked list.

The **Robust** task offered monolingual, bilingual and multilingual tasks using the test collections built over three years: CLEF 2001 - 2003, for six languages: Dutch, English, French, German, Italian and Spanish. Using topics from three years meant that more extensive experiments and a better analysis of the results were possible. The aim of this task was to study and achieve good performance on queries that had proved difficult in the past rather than obtain a high average performance when calculated over all queries.

In this paper we describe the track setup, the evaluation methodology and the participation in the different tasks (Section 2), present the main characteristics of the experiments and show the results (Sections 3 - 5). The final section provides a brief summing up. For information on the various approaches and resources used by the groups participating in this track and the issues they focused on, we refer the reader to the other papers in the Ad Hoc section of these Proceedings.

2 Track Setup

The ad hoc track in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments in the late 1960s. The test collection used consists of a set of “topics” describing information needs and a collection of documents to be searched to find those documents that satisfy these information needs. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures. The distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted to satisfy the particular requirements of the multilingual context. All language dependent tasks such as topic creation and relevance judgment are performed in a distributed setting by native speakers. Rules are established and a tight central coordination is maintained in order to ensure consistency and coherency of topic and relevance judgment sets over the different collections, languages and tracks.

2.1 Test Collections

Different test collections were used in the ad hoc task in 2006. The main (i.e. non-robust) monolingual and bilingual tasks used the same document collections as in Ad Hoc 2005 but new topics were created and new relevance assessments made. As has already been stated, the test collection used for the robust task was derived from the test collections previously developed at CLEF. No new relevance assessments were performed for this task.

Documents. The document collections used for the CLEF 2006 ad hoc tasks are part of the CLEF multilingual corpus of newspaper and news agency documents

Table 1. Document collections for the main stream Ad Hoc tasks

Language	Collections
Bulgarian	Sega 2002, Standart 2002
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94/95
Hungarian	Magyar Hirlap 2002
Portuguese	Público 94/95; Folha 94/95

Table 2. Document collections for the Robust task

Language	Collections
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94
Italian	La Stampa 94, AGZ (SDA) 94/95
Dutch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
German	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanish	EFE 94/95

described in the Introduction to these Proceedings. The Bulgarian and Hungarian collections used in these tasks were new in CLEF 2005 and consist of national newspapers for the year 2002¹. This has meant using collections of different time periods for the ad-hoc mono- and bilingual tasks. This had important consequences on topic creation. Table 1 shows the collections used for each language.

The robust task used test collections containing data in six languages (Dutch, English, German, French, Italian and Spanish) used at CLEF 2001, CLEF 2002 and CLEF 2003. There are approximately 1.35 million documents and 3.6 gigabytes of text in the CLEF 2006 "robust" collection. Table 2 shows the collections used for each language.

Topics. Sets of 50 topics were created for the CLEF 2006 ad hoc mono- and bilingual tasks. One of the decisions taken early on in the organization of the CLEF ad hoc tracks was that the same set of topics would be used to query all collections, whatever the task. There were a number of reasons for this: it makes it easier to compare results over different collections, it means that there is a single master set that is rendered in all query languages, and a single set of relevance assessments for each language is sufficient for all tasks. However, in CLEF 2005 the assessors found that the fact that the collections used in the CLEF 2006 ad hoc mono- and bilingual tasks were from two different time periods (1994-1995 and 2002) made topic creation particularly difficult. It was not possible to create time-dependent topics that referred to particular date-specific events as all topics had to refer to events

¹ It proved impossible to find national newspapers in electronic form for 1994 and/or 1995 in these languages.

that could have been reported in any of the collections, regardless of the dates. This meant that the CLEF 2005 topic set is somewhat different from the sets of previous years as the topics all tend to be of broad coverage. In fact, it was difficult to construct topics that would find a limited number of relevant documents in each collection, and consequently a - probably excessive - number of topics used for the 2005 mono- and bilingual tasks have a very large number of relevant documents.

For this reason, we decided to create separate topic sets for the two different time-periods for the CLEF 2006 ad hoc mono- and bilingual tasks. We thus created two overlapping topic sets, with a common set of time independent topics and sets of time-specific topics. 25 topics were common to both sets while 25 topics were collection-specific, as follows:

- Topics C301 - C325 were used for all target collections
- Topics C326 - C350 were created specifically for the English, French and Portuguese collections (1994/1995)
- Topics C351 - C375 were created specifically for the Bulgarian and Hungarian collections (2002).

This meant that a total of 75 topics were prepared in many different languages (European and non-European): Bulgarian, English, French, German, Hungarian, Italian, Portuguese, and Spanish plus Amharic, Chinese, Hindi, Indonesian, Oromo and Telugu. Participants had to select the necessary topic set according to the target collection to be used.

Below we give an example of the English version of a typical CLEF topic:

```
<top> <num> C302 </num>
<EN-title> Consumer Boycotts </EN-title> <
EN-desc> Find documents that describe or discuss the impact of consumer
boycotts. </EN-desc>
<EN-narr> Relevant documents will report discussions or points of view on
the efficacy of consumer boycotts. The moral issues involved in such
boycotts are also of relevance. Only consumer boycotts are relevant,
political boycotts must be ignored. </EN-narr> </top>
```

For the robust task, the topic sets used in CLEF 2001, CLEF 2002 and CLEF 2003 were used for evaluation. A total of 160 topics were collected and split into two sets: 60 topics used to train the system, and 100 topics used for the evaluation. Topics were available in the languages of the target collections: English, German, French, Spanish, Italian, Dutch.

2.2 Participation Guidelines

To carry out the retrieval tasks of the CLEF campaign, systems have to build supporting data structures. Allowable data structures include any new structures built automatically (such as inverted files, thesauri, conceptual networks, etc.) or manually (such as thesauri, synonym lists, knowledge bases, rules, etc.) from the documents. They may not, however, be modified in response to the topics,

e.g. by adding topic words that are not already in the dictionaries used by their systems in order to extend coverage.

Some CLEF data collections contain manually assigned, controlled or uncontrolled index terms. The use of such terms has been limited to specific experiments that have to be declared as “manual” runs.

Topics can be converted into queries that a system can execute in many different ways. CLEF strongly encourages groups to determine what constitutes a base run for their experiments and to include these runs (officially or unofficially) to allow useful interpretations of the results. Unofficial runs are those not submitted to CLEF but evaluated using the `trec_eval` package. This year we have used the new package written by Chris Buckley for the *Text REtrieval Conference (TREC)* (`trec_eval` 8.0) and available from the TREC website.

As a consequence of limited evaluation resources, we set a maximum number of runs for each task with restrictions on the number of runs that could be accepted for a single language or language combination - we try to encourage diversity.

2.3 Relevance Assessment

The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall values are calculated using pooling techniques. The results submitted by the groups participating in the ad hoc tasks are used to form a pool of documents for each topic and language by collecting the highly ranked documents from all submissions. This pool is then used for subsequent relevance judgments. The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [1] with respect to the CLEF 2003 pools. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed. New pools were formed in CLEF 2006 for the runs submitted for the main stream mono- and bilingual tasks and the relevance assessments were performed by native speakers. Instead, the robust tasks used the original pools and relevance assessments from CLEF 2001-2003.

The individual results for all official ad hoc experiments in CLEF 2006 are given in the Appendix at the end of the on-line Working Notes prepared for the Workshop [2] and available online at www.clef-campaign.org.

2.4 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of *Information Retrieval Systems (IRSs)* can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participants and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [3]. For the robust task, we used different measures, see below Section 5.

2.5 Participants and Experiments

A total of 25 groups from 15 different countries submitted results for one or more of the ad hoc tasks - a slight increase on the 23 participants of last year.

A total of 296 experiments were submitted with an increase of 16% on the 254 experiments of 2005. On the other hand, the average number of submitted runs per participant is nearly the same: from 11 runs/participant of 2005 to 11.7 runs/participant of this year.

Participants were required to submit at least one title+description (“TD”) run per task in order to increase comparability between experiments. The large majority of runs (172 out of 296, 58.11%) used this combination of topic fields, 78 (26.35%) used all fields, 41 (13.85%) used the title field, and only 5 (1.69%)

Table 3. Breakdown of experiments into tracks and topic languages

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
Monolingual-BG	4	11
Monolingual-FR	8	27
Monolingual-HU	6	17
Monolingual-PT	12	37
Bilingual-X2BG	1	2
Bilingual-X2EN	5	33
Bilingual-X2FR	4	12
Bilingual-X2HU	1	2
Bilingual-X2PT	6	22
Robust-Mono-DE	3	7
Robust-Mono-EN	6	13
Robust-Mono-ES	5	11
Robust-Mono-FR	7	18
Robust-Mono-IT	5	11
Robust-Mono-NL	3	7
Robust-Bili-X2DE	2	5
Robust-Bili-X2ES	3	8
Robust-Bili-X2NL	1	4
Robust-Multi	4	10
Robust-Training-Mono-DE	2	3
Robust-Training-Mono-EN	4	7
Robust-Training-Mono-ES	3	5
Robust-Training-Mono-FR	5	10
Robust-Training-Mono-IT	3	5
Robust-Training-Mono-NL	2	3
Robust-Training-Bili-X2DE	1	1
Robust-Training-Bili-X2ES	1	2
Robust-Training-Multi	2	3
Total		296

(b) List of experiments by topic language.

Topic Lang.	# Runs
English	65
French	60
Italian	38
Portuguese	37
Spanish	25
Hungarian	17
German	12
Bulgarian	11
Indonesian	10
Dutch	10
Amharic	4
Oromo	3
Hindi	2
Telugu	2
Total	296

used the description field. The majority of experiments were conducted using automatic query construction (287 out of 296, 96.96%) and only in a small fraction of the experiments (9 out 296, 3.04%) have queries which were manually constructed from topics. A breakdown into the separate tasks is shown in Table 3(a).

Fourteen different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with French second. The number of runs per topic language is shown in Table 3(b).

3 Main Stream Monolingual Experiments

Monolingual retrieval was offered for Bulgarian, French, Hungarian, and Portuguese. As can be seen from Table 3(a), the number of participants and runs for each language was quite similar, with the exception of Bulgarian, which had a slightly smaller participation. This year just 6 groups out of 16 (37.5%) submitted monolingual runs only (down from ten groups last year), and 5 of these groups were first time participants in CLEF. Most of the groups submitting monolingual runs were doing this as part of their bilingual or multilingual system testing activity. Details on the different approaches used can be found in the papers in this section of the Proceedings. There was a lot of detailed work with Portuguese language processing; not surprising as we had four new groups from Brazil in Ad Hoc this year. As usual, there was a lot of work on the development of stemmers and morphological analysers ([4], for instance, applies a very deep morphological analysis for Hungarian) and comparisons of the pros and cons of so-called "light" and "heavy" stemming approaches (e.g. [5]). In contrast to previous years, we note that a number of groups experimented with NLP techniques (see, for example, papers by [6], and [7]).

Table 4. Best entries for the monolingual track

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Bulgarian	unine	rsi-jhu	hummingbird	daedalus		1st vs 4th
MAP	33.14%	31.98%	30.47%	27.87%		20.90%
Run	UniNEbg2	02aplmobgtd4	humBG06tde	bgFSbg2S		
French	unine	rsi-jhu	hummingbird	alicante	daedalus	1st vs 5th
MAP	44.68%	40.96%	40.77%	38.28%	37.94%	17.76%
Run	UniNEfr3	95aplmofrtd5s1	humFR06tde	8dfrexp	frFSfr2S	
Hungarian	unine	rsi-jhu	alicante	mokk	hummingbird	1st vs 5th
MAP	41.35%	39.11%	35.32%	34.95%	32.24%	28.26%
Run	UniNEhu2	02aplmohutd4	30dfrexp	plain2	humHU06tde	
Portuguese	unine	hummingbird	alicante	rsi-jhu	u.buffalo	1st vs 5th
MAP	45.52%	45.07%	43.08%	42.42%	40.53%	12.31%
Run	UniNEpt1	humPT06tde	30okapiexp	95aplmopttd5	UBptTDrfl	

3.1 Results

Table 4 shows the results for the top five groups for each target collection, ordered by mean average precision. The table reports: the short name of the participating group; the mean average precision achieved by the run; the run identifier; and the performance difference between the first and the last participant. Table 4 regards runs using title + description fields only (the mandatory run).

4 Main Stream Bilingual Experiments

The bilingual task was structured in four subtasks ($X \rightarrow \text{BG, FR, HU or PT}$ target collection) plus, as usual, an additional subtask with English as target language restricted to newcomers in a CLEF cross-language task. This year, in this subtask, we focussed in particular on non-European topic languages and in particular languages for which there are still few processing tools or resources in existence. We thus offered two Ethiopian languages: Amharic and Oromo; two Indian languages: Hindi and Telugu; and Indonesian. Although, as was to be expected, the results are not particularly good, we feel that experiments of this type with lesser-studied languages are very important (see papers by [8], [9])

4.1 Results

Table 5 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision. Again both pooled and not pooled runs are included in the best entries for each track, with the exception of Bilingual $X \rightarrow \text{EN}$.

For bilingual retrieval evaluation, a common method to evaluate performance is to compare results against monolingual baselines. For the best bilingual systems, we have the following results for CLEF 2006:

- $X \rightarrow \text{BG}$: 52.49% of best monolingual Bulgarian IR system;

Table 5. Best entries for the bilingual task

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Bulgarian MAP Run	daedalus 17.39% bgFSBgWen2S					
French MAP Run	unine 41.92% UniNEBifr1	queenmary 33.96% QMUL06e2f10b	rsi-jhu 33.60% aplbienfrd	daedalus 33.20% frFSfrSen2S		1st vs 4th 26.27%
Hungarian MAP Run	daedalus 21.97% huFShuMen2S					
Portuguese MAP Run	unine 41.38% UniNEBipt2	rsi-jhu 35.49% aplbiesptd	queenmary 35.26% QMUL06e2p10b	u.buffalo 29.08% UBen2ptTDrf2	daedalus 26.50% ptFSptSen2S	1st vs 5th 55.85%
English MAP Run	rsi-jhu 32.57% aplbiiin5	depok 26.71% UItd_mt	ltrc 25.04% OMTD	celi 23.97% CELItitleNOEXPANSION	dsv 22.78% DsvAmhEngFullNofuzz	1st vs 5th 42.98%

- $X \rightarrow \text{FR}$: 93.82% of best monolingual French IR system;
- $X \rightarrow \text{HU}$: 53.13% of best monolingual Hungarian IR system.
- $X \rightarrow \text{PT}$: 90.91% of best monolingual Portuguese IR system;

We can compare these to those for CLEF 2005:

- $X \rightarrow \text{BG}$: 85% of best monolingual Bulgarian IR system;
- $X \rightarrow \text{FR}$: 85% of best monolingual French IR system;
- $X \rightarrow \text{HU}$: 73% of best monolingual Hungarian IR system.
- $X \rightarrow \text{PT}$: 88% of best monolingual Portuguese IR system;

While these results are good for the well-established-in-CLEF languages, and can be read as state-of-the-art for this kind of retrieval system, at a first glance they appear disappointing for Bulgarian and Hungarian. However, we must point out that, unfortunately, this year only one group submitted cross-language runs for Bulgarian and Hungarian and thus it does not make much sense to draw any conclusions from these, apparently poor, results for these languages. It is interesting to note that when *Cross Language Information Retrieval (CLIR)* system evaluation began in 1997 at TREC-6 the best CLIR systems had the following results:

- $\text{EN} \rightarrow \text{FR}$: 49% of best monolingual French IR system;
- $\text{EN} \rightarrow \text{DE}$: 64% of best monolingual German IR system.

5 Robust Experiments

The robust task was organized for the first time at CLEF 2006. The evaluation of robustness emphasizes stable performance over all topics instead of high average performance [10]. The perspective of each individual user of an information retrieval system is different from the perspective of an evaluation initiative. Users are disappointed by systems which deliver poor results for some topics whereas an evaluation initiative rewards systems which deliver good average results. A system delivering poor results for hard topics is likely to be considered of low quality by a user although it may still reach high average results. The robust task has been inspired by the robust track at TREC where it was organized at TREC 2003, 2004 and 2005. A robust evaluation stresses performance for weak topics. This can be achieved by employing the *Geometric Average Precision (GMAP)* as a main indicator for performance instead of the *Mean Average Precision (MAP)* of all topics. Geometric average has proven to be a stable measure for robustness at TREC [10]. The robust task at CLEF 2006 is concerned with multilingual robustness. It is essentially an ad-hoc task which offers mono-lingual and cross-lingual sub tasks.

As stated, the robust task used test collections developed in CLEF 2001, CLEF 2002 and CLEF 2003. No additional relevance judgements were made this year for this task. However, the data collection was not completely constant over all three CLEF campaigns which led to an inconsistency between

relevance judgements and documents. The SDA 95 collection has no relevance judgements for most topics (#41 - #140). This inconsistency was tolerated in order to increase the size of the collection. One participant reported that exploiting the knowledge would have resulted in an increase of approximately 10% in MAP [11]. However, participants were not allowed to use this information. The results of the original submissions for the data sets were analyzed in order to identify the most difficult topics. This turned out to be a very hard task. The difficulty of a topic varies greatly among languages, target collections and tasks. This confirms the finding of the TREC 2005 robust task where the topic difficulty differed greatly even for two different English collections. Topics are not inherently difficult but only in combination with a specific collection [12]. Topic difficulty is usually defined by low MAP values for a topic. We also considered a low number of relevant documents and high variation between systems as indicators for difficulty. Because no consistent definition of topic difficulty could be found, the topic set for the robust task at CLEF 2006 was arbitrarily split into two sets. Participants were allowed to use the available relevance assessments for the set of 60 training topics. The remaining 100 topics formed the test set for which results are reported. The participants were encouraged to submit results for training topics as well. These runs will be used to further analyze topic difficulty.

The robust task received a total of 133 runs from eight groups. Most popular among the participants were the mono-lingual French and English tasks. For the multi-lingual task, four groups submitted ten runs. The bi-lingual tasks received fewer runs. A run using title and description was mandatory for each group. Participants were encouraged to run their systems with the same setup for all robust tasks in which they participated (except for language specific resources). This way, the robustness of a system across languages could be explored.

Effectiveness scores for the submissions were calculated with the GMAP which is calculated as the n -th root of a product of n values. GMAP was computed using the version 8.0 of `trec_eval`² program. In order to avoid undefined result figures, all precision scores lower than 0.00001 are set to 0.00001.

5.1 Robust Monolingual Results

Table 6 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

Hummingbird submitted the best results for five out of six sub tasks. However, the differences between the best runs are small and not always statistically significant, see [21,2].

The MAP figures were above 45% for five out of six sub tasks. These numbers can be considered as state of the art.

It is striking that the rankings based on the MAP is identical to the ranking based on the GMAP measure in most cases.

² http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz

Table 6. Best entries for the robust monolingual task

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Dutch	hummingbird	daedalus	colesir			1st vs 3rd
MAP	51.06%	42.39%	41.60%			22.74%
GMAP	25.76%	17.57%	16.40%			57.13%
Run	humNL06Rtde	nlFSnlR2S	CoLesIRnlTst			
English	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	47.63%	43.66%	43.48%	39.69%	37.64%	26.54%
GMAP	11.69%	10.53%	10.11%	8.93%	8.41%	39.00%
Run	humEN06Rtde	reinaENTdtest	dcudesceng12075	enFSenR2S	CoLesIRenTst	
French	unine	hummingbird	reina	dcu	colesir	1st vs 5th
MAP	47.57%	45.43%	44.58%	41.08%	39.51%	20.40%
GMAP	15.02%	14.90%	14.32%	12.00%	11.91%	26.11%
Run	UniNEfrr1	humFR06Rtde	reinaFRtdtest	dcudescfr12075	CoLesIRfrTst	
German	hummingbird	colesir	daedalus			1st vs 3rd
MAP	48.30%	37.21%	34.06%			41.81%
GMAP	22.53%	14.80%	10.61%			112.35%
Run	humDE06Rtde	CoLesIRdeTst	deFSdeR2S			
Italian	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	41.94%	38.45%	37.73%	35.11%	32.23%	30.13%
GMAP	11.47%	10.55%	9.19%	10.50%	8.23%	39.37%
Run	humIT06Rtde	reinaITtdtest	dcudescit1005	itFSitR2S	CoLesIRitTst	
Spanish	hummingbird	reina	dcu	daedalus	colesir	1st vs 5th
MAP	45.66%	44.01%	42.14%	40.40%	40.17%	13.67%
GMAP	23.61%	22.65%	21.32%	19.64%	18.84%	25.32%
Run	humES06Rtde	reinaEStdtest	dcudescsp12075	esFSesR2S	CoLesIResTst	

5.2 Robust Bilingual Results

Table 7 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision).

As stated in 4.1, for bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2006:

- $X \rightarrow DE$: 60.37% of best monolingual German IR system;
- $X \rightarrow ES$: 80.88% of best monolingual Spanish IR system;
- $X \rightarrow NL$: 69.27% of best monolingual Dutch IR system.

5.3 Robust Multilingual Results

Table 8 shows the best results for this task for runs using the title+description topic fields. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). The figures are lower than for multilingual experiments at previous CLEF campaigns. This shows that the multilingual retrieval problem is far from being solved and that results depend much on the topic set.

Table 7. Best entries for the robust bilingual task.

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Dutch	daedalus					
MAP	35.37%					
GMAP	9.75%					
Run	nIFSnlRLfr2S					
German	daedalus	colesir				1st vs 2nd
MAP	29.16%	25.24%				15.53%
GMAP	5.18%	4.31%				20.19%
Run	deFSdeRSen2S	CoLesIRendeTst				
Spanish	reina	dcu	daedalus			1st vs 3rd
MAP	36.93%	33.22%	26.89%			37.34%
GMAP	13.42%	10.44%	6.19%			116.80%
Run	reinaIT2ESStdtest	dcuitqydescsp12075	esFSesRLit2S			

Table 8. Best entries for the robust multilingual task

Track	Participant Rank					Diff.
	1st	2nd	3rd	4th	5th	
Multilingual	jaen	daedalus	colesir	reina		1st vs 4th
MAP	27.85%	22.67%	22.63%	19.96%		39.53%
GMAP	15.69%	11.04%	11.24%	13.25%		18.42%
Run	ujamlrsv2	mlRSFSen2S	CoLesIRmultTst	reinaES2mtdtest		

5.4 Comments on Robust Cross Language Experiments

The robust track is especially concerned with the performance for hard topics which achieve low MAP figures. One important reason for weak topics is the lack of good keywords in the query and difficulties to expand the query properly within the collection. A strategy often applied is the query expansion with external collections like the web. This or other strategies are sometimes applied depending on the topic. Only when a topic is classified as a difficult topic, additional techniques are applied. Several participants relied on the high correlation between the measure and optimized their systems as in previous campaigns. Nevertheless, some groups worked specifically for robustness. The SINAI system took an approach which has proved successful at the TREC robust task, expansion with terms gathered from a web search engine [13]. The REINA system from the University of Salamanca used a heuristic to determine hard topics during training. Subsequently, different expansion techniques were applied [14]. The MIRACLE system tried to find a fusion scheme which had a positive effect on the robust measure [16]. The results are mixed. Savoy & Abdou reported that expansion with an external search engine did not improve the results [11]. It seems that optimal heuristics for the selection of good expansion terms still

need to be developed. Hummingbird thoroughly discussed alternative evaluation measures for capturing the robustness of runs. [15].

6 Conclusions

We have reported the results of the ad hoc cross-language textual document retrieval track at CLEF 2006. This track is considered to be central to CLEF as for many groups it is the first track in which they participate and provides them with an opportunity to test their systems and compare performance between monolingual and cross-language runs, before perhaps moving on to more complex system development and subsequent evaluation. However, the track is certainly not just aimed at beginners. It also gives groups the possibility to measure advances in system performance over time. In addition, each year, we also include a task aimed at examining particular aspects of cross-language text retrieval. This year, the focus was examining the impact of "hard" topics on performance in the "robust" task.

Thus, although the ad hoc track in CLEF 2006 offered the same target languages for the main mono- and bilingual tasks as in 2005, it also had two new focuses. Groups were encouraged to use non-European languages as topic languages in the bilingual task. We were particularly interested in languages for which few processing tools were readily available, such as Amharic, Oromo and Telugu. In addition, we set up the "robust task" with the objective of providing the more expert groups with the chance to do in-depth failure analysis.

For reasons of space, in this paper we have only been able to summarise the main results; more details, including sets of statistical analyses can be found in [21,2].

Finally, it should be remembered that, although over the years we vary the topic and target languages offered in the track, all participating groups also have the possibility of accessing and using the test collections that have been created in previous years for all of the twelve languages included in the CLEF multilingual test collection. The test collections for CLEF 2000 - CLEF 2003 are about to be made publicly available on the *Evaluations and Language resources Distribution Agency (ELDA)* catalog³.

References

1. Braschler, M.: CLEF 2003 - Overview of results. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 44-63. Springer, Heidelberg (2004)
2. Di Nunzio, G.M., Ferro, N.: Appendix A. Results of the Core Tracks. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Working Notes for the CLEF 2006 Workshop (2006), Published Online at www.clef-campaign.org
3. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 7-20. Springer, Heidelberg (2004)

³ <http://www.elda.org/>

4. Halácsy, P., Trón, V.: Benefits of Deep NLP-based Lemmatization for Information Retrieval. [In this volume]
5. Moreira Orenge, V., Buriol, L.S., Ramos Coelho, A.: A Study on the use of Stemming for Monolingual Ad-Hoc Portuguese. *Information Retrieval* (2006)
6. Azevedo Arcoverde, J.M., das Gracias Volpe Nunes, M.: NLP-Driven Constructive Learning for Filtering an IR Document Stream. [In this volume]
7. Gonzalez, M., de Lima, V.L.S.: The PUCRS-PLN Group participation at CLEF 2006. [In this volume]
8. Pingali, P., Tune, K.K., Varma, V.: Hindi, Telugu, Oromo, English CLIR Evaluation. [In this volume]
9. Hayurani, H., Sari, S., Adriani, M.: Query and Document Translation for English-Indonesian Cross Language IR. [In this volume]
10. Voorhees, E.M.: The TREC Robust Retrieval Track. *SIGIR Forum* 39, 11–20 (2005)
11. Savoy, J., Abdou, S.: Experiments with Monolingual, Bilingual, and Robust Retrieval. [In this volume]
12. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In: Voorhees, E.M., Buckland, L.P. (eds.): *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)* [last visited 2006, August 4] (2005), http://trec.nist.gov/pubs/trec14/t14_proceedings.html
13. Martinez-Santiago, F., Montejo-Ráez, A., Garcia-Cumbreras, M., Ureña-Lopez, A.: SINAI at CLEF 2006 Ad-hoc Robust Multilingual Track: Query Expansion using the Google Search Engine. [In this volume] (2006)
14. Zazo, A., Berrocal, J., Figuerola, C.: Local Query Expansion Using Term Windows for Robust Retrieval. [In this volume]
15. Tomlinson, S.: Comparing the Robustness of Expansion Techniques and Retrieval Measures. [In this volume]
16. Goni-Menoyo, J., Gonzalez-Cristobal, J., Vilena-Román, J.: Report of the MIRA-CLE teach for the Ad-hoc track in CLEF 2006. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) *Working Notes for the CLEF 2006 Workshop*, Published Online (2006)
17. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Korfhage, R., Rasmussen, E., Willett, P. (eds.) *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*, pp. 329–338. ACM Press, New York (1993)
18. Conover, W.J.: *Practical Nonparametric Statistics*, 1st edn. John Wiley and Sons, New York (1971)
19. Judge, G.G., Hill, R.C., Griffiths, W.E., Lütkepohl, H., Lee, T.C.: *Introduction to the Theory and Practice of Econometrics*, 2nd edn. John Wiley and Sons, New York (1988)
20. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: Sparck Jones, K., Willett, P. (eds.) *Readings in Information Retrieval*, pp. 205–216. Morgan Kaufmann Publisher, Inc, San Francisco, California (1997)
21. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In: Nardi, A., Peters, C., Vicedo, J.L., eds.: *Working Notes for the CLEF 2006 Workshop* (2006) (last visited, March 23, 2007), http://www.clef-campaign.org/2006/working_notes/workingnotes2006/dinunzio0CLEF2006.pdf

Hindi, Telugu, Oromo, English CLIR Evaluation

Prasad Pingali, Kula Kekeba Tune, and Vasudeva Varma

Language Technologies Research Centre,
IIIT, Hyderabad, India
{pvvpr, vv}@iiit.ac.in
{kulakk}@students.iiit.ac.in
<http://search.iiit.ac.in>

Abstract. This paper presents the Cross Language Information Retrieval (CLIR) experiments of Language Technologies Research Centre (LTRC, IIIT-Hyderabad) as part of our participation in the ad-hoc track of CLEF 2006. This is our first participation in the CLEF evaluation campaign and we focused on Afaan Oromo, Hindi and Telugu as source (query) languages for retrieval of documents from English text collection. We have used a dictionary based approach for CLIR. After a brief description of our CLIR system we discuss the evaluation results of various experiments we conducted using CLEF 2006 dataset.

Keywords: Hindi-English, Telugu-English, Oromo-English, CLIR experiments, CLEF 2006.

1 Introduction

Cross-Language Information Retrieval (CLIR) can briefly be defined as a subfield of Information Retrieval (IR) system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query. Thus, CLIR research mainly deals with the study of IR systems that accept queries (or information needs) in one language and return objects of a different language. These objects could be text documents, passages, images, audio or video documents. Some of the key technical issues [1] for cross language information retrieval can be thought of as:

- How can a query term in L_1 be expressed in L_2 ?
- What mechanisms determine which of the possible translations of text from L_1 to L_2 should be retained?
- In cases where more than one translation are retained, how can different translation alternatives be weighed?

In order to address these issues, many different techniques were tried in various CLIR systems in the past [2]. In this paper we present and discuss our dictionary-based bilingual information retrieval experiments that had been conducted at CLEF 2006 by our CLIR research group from Language Technologies Research Centre (LTRC) of IIIT-Hyderabad, India. We had participated in CLEF ad hoc track for the first time by conducting seven various bilingual

information retrieval experiments (official runs) for three language pairs: Afaan Oromo, Hindi and Telugu to English. Our main purpose was to obtain hands-on experience in joint evaluations of CLIR forum by conducting experiments in two major Indian languages (i.e. Hindi and Telugu) and one major Ethiopian language (i.e. Afaan Oromo) as source languages for retrieval of documents from a large size English test collection. In particular, we want to investigate and assess the performance level that we could achieve by combining and using the scarcely available resources of these languages. All of our CLIR experiments were conducted by using dictionary-based query translation techniques based on the CLEF 2006 experimental setup.

2 Review of Related Works

Very little work has been done in the past in the areas of IR and CLIR involving Indian and African languages. In relation to Indian languages, a surprise language exercise [3] was conducted at ACM TALIP¹ in 2003. The task was to build CLIR systems for English to Hindi and Cebuano, where the queries were in English and the documents were in Hindi and Cebuano. Five teams participated in this evaluation task at ACM TALIP providing some insights into the issues involved in processing Indian language content. A few other information access systems were built apart from this task such as cross language Hindi headline generation [4], English to Hindi question answering system [5] etc. We previously built a monolingual web search engine for various Indian languages which is capable of retrieving information from multiple character encodings [6]. However, no work was found related to CLIR involving Telugu or any other Indian language other than Hindi.

Like wise, very few CLIR research works have been conducted in relation to African indigenous languages including major Ethiopian languages. A case study for Zulu (one of the major languages in South Africa) was reported by [7] in relation to application for cross-lingual information access to knowledge databases. Another similar study was undertaken by [8] on Afrikaans-English cross-language information retrieval. The main components of this CLIR were source and target language normalizers, a translation dictionary, source and target language stopword lists and an approximate string matching. A dictionary-based query translation technique was used to translate Afrikaans queries into English queries in similar manner to our approach. The performance of the CLIR system was evaluated by using 35 topics from the CLEF 2001 English test collection (employing title and descriptions fields of the topic sets) and achieved an average precision 19.4%.

More recently, different dictionary-based Amharic-English and Amharic-French CLIR experiments were conducted at a series of CLEF ad hoc tracks [9,10,11]. The first Amharic-English information retrieval was conducted at CLEF 2004 with emphasis on analyzing the impact of the stopword lists of

¹ ACM Transactions on Asian Language Information Processing.
<http://www.acm.org/pubs/talip/>

the source and target languages. After a year, similar Amharic-French CLIR experiment were conducted by using two different search engines and reported at CLEF 2005. More recently, another dictionary based Amharic-English CLIR experiment was carried out at CLEF 2006 by employing Amharic morphological analyzer and part of speech tagging. A mean average precision of 22.78% was obtained and reported in this recent Amharic-English CLIR experiment [11]. On the other hand, to the best of the current researchers knowledge no formal study has been reported in relation to CLIR involving Afaan Oromo or a number of other Ethiopian Languages except Amharic.

Furthermore, some research was previously done in the areas of Machine Translation (MT) involving Indian languages [12]. Most of the Indian language MT efforts involve studies on translating various Indian languages amongst themselves or translating English into Indian language content. Hence most of the Indian language resources available for our work are largely biased to these tasks. This led to the challenge of using resources which enabled translation from English to Indian languages for a task involving translation from Indian languages to English.

3 Our System

We have used dictionary based query translation methods for all of our bilingual retrieval experiments. The ranking is achieved using a vector based retrieval and ranking model using TFIDF ranking algorithm [13]. In our post-CLEF experiments we have used probabilistic transliteration and language modeling approaches as part of our system.

3.1 Query Translation

We used Hindi-English, Telugu-English and Oromo-English bilingual dictionaries to obtain word-word translations for query translation. We then perform a lemmatization process in order to obtain the stems of the given input queries. The details of the lemmatization process Hindi, Telugu and Oromo are similar to those mentioned in [14,15,16] respectively.

The terms remaining after suffix removal are looked up in bilingual dictionary. A set of multiple English meanings for a given query term would be obtained for a given source language term. Many of the terms may not be found in the bilingual lexicon since the term is a proper name or a word from a foreign language or a valid source language word which just did not occur in the dictionary. In some of the cases dictionary lookup for a term might also fail because of improper stemming or suffix removal. All the lookup failure cases are attempted for automatic transliteration. For Afaan Oromo, we made use of a named entity list which was available both in English and Oromo to lookup transliterations. However, for Indian language queries our official submission was based on a phonetic transliteration algorithm. After the CLEF 2006 official task, we have used a probabilistic transliteration model instead of a phonetic algorithm which is described in the next section.

3.2 Query Transliteration

In order to obtain automatic transliterations of the source language query terms, we used phonetic and probabilistic techniques as described below.

Phonetic Transliteration. The transliteration was first performed with a set of phoneme mappings between Indian language and English. While this technique might succeed in a few cases, in many cases this may not transliterate into the right English term. Therefore we used approximate string matching algorithms of the obtained transliteration against the lexicon from the corpus. We used the double metaphone [17] algorithm as well as Levensteins approximate string matching algorithm to obtain possible transliterations for the query term which was not found in the dictionary. The intersection set from these two algorithms were added to the translated English query terms.

Probabilistic Transliteration. We use a probabilistic transliteration model for Indian languages which was previously suggested in [14] for Arabic-English language pair. For obtaining multiple transliterations with ranking for a source language word w_s into target language as w_t , where A is the vocabulary of the target language and t_i and s_i are the i^{th} characters in the target and source languages respectively is given by,

$$\begin{aligned} P(w_t|w_s, w_t \in A) &= P(w_t|w_s) \cdot P(w_t \in A) \\ \text{where, } P(w_t|w_s) &= \prod_i P(t_i|s_i) \\ \text{and, } P(w_t \in A) &= \prod_i P(t_i|s_{i-1}) \end{aligned} \tag{1}$$

We have used a parallel transliterated word list of 30,000 unique words to generate the model for Hindi-English and Telugu-English transliteration using GIZA++ software, and achieved transliteration of a source language query term using the equation 1.

3.3 Query Refinement and Retrieval

For our Indian language to English experiments we used a query refinement algorithm in order to minimize or eliminate noisy terms in the translated query. In our official submission, we achieved this task using a pseudo-relevance feedback technique discussed in the next section. In our Post-CLEF experiments we used a language modeling approach which showed improvements in our CLIR system performance. We discuss the evaluation of our Indian language to English CLIR in section 4.

Using Pseudo-Relevance Feedback. Once the translation and transliteration tasks are performed on the input Hindi and Telugu queries, we tried to

address the second issue for CLIR from our list as mentioned in section 1. We tried to prune out the possible translations for the query in an effort to reduce the possible noise in translations. In order to achieve this, we used a pseudo-relevance feedback based on the top ranking documents above a threshold using the TFIDF retrieval engine. The translated English query was issued to the lucene search engine and a set of top 10 documents were retrieved. The translated English terms that did not occur in these documents were pruned out in an effort to reduce the noisy translations.

Language Modeling Approach. It was shown that language modeling frameworks perform better than the vector space models such as TFIDF algorithms. With the same motivation we applied the language modeling technique as described in [18] to retrieve and rank documents after translation and transliteration of the source language query. We compute the relevance score of a document to be the probability that a given document’s language model emits query terms and does not emit other terms. Therefore,

$$R(Q, D) = P(Q|D) \approx \prod_{t \in Q} P(t|M_d) \cdot \prod_{t \notin Q} (1 - P(t|M_d)) \quad (2)$$

where Q is the query and t is every term emitted by the document’s language model and M_d is the document’s language model.

We then extended the language modeling framework to include post-translation query expansion achieved by the joint probability of cooccurrence of terms from the target language corpus. We rank documents with expanded features using the equation 3.

$$R(Q, D) = P(Q|D) \approx \prod_{t_i \in D} \prod_{t_j \in Q} P(t_j, t_i) \cdot P(t_i|M_d) \cdot \prod_{t_j \notin Q} (1 - P(t_j, t_i)) \cdot P(t_i|M_d) \quad (3)$$

where $P(t_j, t_i)$ is the joint probability of the two terms t_i and t_j cooccurring together in a window of length k in the corpus. If k is equal to 0, such a probability distribution would result in bigram probabilities. We used the window length as 8 which has been experimentally shown to capture semantically related words in [19].

4 Results and Discussions

We submitted seven official runs (bilingual experiments) that differed in terms of the source languages and the utilized fields in the topic set and conducted two post-CLEF experimental runs. A summary of these runs together with their run labels (ids), source languages and topic fields employed for our experiments are given in Table 1.

Table 1. Summary of the seven official runs and two post-CLEF runs

Run-Id	Source Language	Run Description
HNT	Hindi	Title run
HNTD	Hindi	Title and description run
TET	Telugu	Title run
TETD	Telugu	Title and description run
OMT	Afaan Oromo	Title run
OMTD	Afaan Oromo	Title and description run
OMTDN	Afaan Oromo	Title, description and narration run
HNLM	Hindi	Title, description run using Language Modeling
TELM	Telugu	Title, description run using Language Modeling

Table 2. Summary of average results for various CLIR runs

Run-Id	Rel-tot.	Rel-Ret.	MAP (%)	R-Prec.(%)	GAP (%)	B-Pref.(%)
OMT	1,258	870	22.00	24.33	7.50	20.60
OMTD	1,258	848	25.04	26.24	9.85	23.52
OMTDN	1,258	892	24.50	25.72	9.82	23.41
HNT	1,258	714	12.32	13.14	2.40	12.78
HNTD	1,258	650	12.52	13.16	2.41	10.91
TET	1,258	565	8.09	8.39	0.34	8.36
TETD	1,258	554	8.16	8.42	0.36	7.84
HNLM	1,258	1051	26.82	26.33	9.41	25.19
TELM	1,258	1018	23.72	25.50	9.17	24.35

The Mean Average Precision (MAP) scores, total number of relevant documents (Rel-tot), the number of retrieved relevant documents (Rel-Ret.), the non-interpolated average precision (R-Prec) and binary preference measures (B-Pref) of the various runs described in table 1 are summarized and presented in Tables 2 and 3.

Table 2 shows that a Mean Average Precision (MAP) of 25.04% that was obtained with our OMTD run (i.e. the title and description run of Afaan Oromo topics) is the highest performance level that we have achieved in our official CLEF 2006 experiments. The overall relatively low performance of the CLIR system with the four official submissions of Indian languages' queries when compared to Afaan Oromo, suggests that a number of Hindi and Telugu topics have low performance statistics, which was also evident from the topic wise score breakups. This is indicative of the fact that our current simple techniques such as dictionary lookup with minimal lemmatization like suffix removal and retrieval and ranking using vector space models may not be sufficient for Indian language CLIR. With these findings in our official CLEF participation, we used probabilistic models for automatic transliteration and retrieval and ranking of documents using language modeling techniques. Our post-CLEF experiments on Hindi and Telugu queries show very good performance with MAP of 26.82% and

Table 3. Interpolated RecallPrecision scores for all runs in %

Recall	OMT	OMTD	OMTDN	HNT	HNTD	TET	TETD	HNLM	TELM
0	48.73	58.01	59.50	35.03	31.83	17.23	18.74	52.80	51.85
10	39.93	47.75	46.45	28.86	27.76	14.53	15.62	46.27	43.65
20	34.94	42.33	37.77	22.99	21.56	12.76	12.80	40.23	34.89
30	30.05	32.15	31.17	15.03	15.63	10.96	10.51	34.33	30.07
40	26.41	28.55	28.27	11.76	12.66	10.07	9.47	31.43	27.57
50	22.98	24.90	24.72	10.62	10.78	9.79	8.79	27.41	24.47
60	18.27	20.19	19.40	8.49	9.32	7.20	7.33	23.21	20.94
70	15.10	16.59	15.61	6.53	7.42	4.78	5.37	21.35	18.08
80	11.76	12.87	12.70	4.90	5.76	3.80	3.94	18.18	15.84
90	8.58	8.37	8.56	3.87	4.44	2.91	2.89	12.80	10.71
100	6.56	6.05	6.58	3.15	3.26	2.19	1.98	8.76	7.06

23.72% respectively. However, we believe that the CLIR performance for Indian language queries can further be improved with an improvement in the stemming algorithm and better coverage in the bilingual dictionaries.

5 Conclusions and Future Directions

In this paper we described our CLIR system for Hindi-English, Telugu-English and Oromo-English language pairs, together with the evaluation results of official and post-CLEF runs conducted using CLEF 2006 evaluation data. Based on our dictionary-based CLIR evaluation experiments for three different languages we attempted to show how very limited language resources can be used in a bilingual information retrieval setting. Since this is the first time we participated in CLEF campaign, we concentrated on evaluation of the over all performance of our experimental CLIR system that is being developed at our research center. We feel we have obtained reasonable average results for some of the official runs, given the limited resources and simple approaches we have used in our experiments. This is very encouraging because there is a growing need for development and application of CLIR systems for a number of Indian and African languages. However, there is still lots of room for improvement in the performance of our CLIR system.

References

1. Grefenstette, G., Grefenstette, G.: Cross-Language Information Retrieval. Kluwer Academic Publishers, Norwell (1998)
2. Oard, D.: Alternative approaches for cross language text retrieval. In: AAAI Symposium on Cross Language Text and Speech Retrieval, USA (1997)
3. Oard, D.W.: The surprise language exercises. ACM Transactions on Asian Language Information Processing (TALIP) 2(2), 79–84 (2003)

4. Dorr, B., Zajic, D., Schwartz, R.: Cross-language headline generation for hindi. *ACM Transactions on Asian Language Information Processing (TALIP)* 2(3), 270–289 (2003)
5. Sekine, S., Grishman, R.: Hindi-English cross-lingual question-answering system. *ACM Transactions on Asian Language Information Processing (TALIP)* 2(3), 181–192 (2003)
6. Pingali, P., Jagarlamudi, J., Varma, V.: Webkhoj: Indian language ir from multiple character encodings. In: *WWW '06: Proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland, pp. 801–809. ACM Press, New York (2006)
7. Cosijn, E., Pirkola, A., Bothma, T., Jrvlin, K.: Information access in indigenous languages: a case study in Zulu. In: *Proceedings of the fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*, Seattle, USA (2002)
8. Cosijn, E., Keskustalo, H., Pirkola, A.: Afrikaans - English Cross-language Information Retrieval. In: *Proceedings of the 3rd biennial DISSAnet Conference*, Pretoria (2004)
9. Alemu, A., Asker, L., Coster, R., Karlgen, J.: Dictionary Based Amharic English Information Retrieval. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004. LNCS*, vol. 3491, Springer, Heidelberg (2005)
10. Alemu, A., Asker, L., Coster, R., Karlgen, J.: Dictionary Based Amharic French Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
11. Alemu, A., Asker, L., Coster, R., Karlgen, J.: Dictionary Based Amharic English Information Retrieval. In: *CLEF 2006, Bilingual Task* (2006)
12. Bharati, A., Sangal, R., Sharma, D.M., Kulkarni, A.P.: Machine translation activities in India: A survey. In: *the Proceedings of workshop on survey on Research and Development of Machine Translation in Asian Countries* (2002)
13. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. *Information Process. Management* 24(5), 513–523 (1988)
14. Larkey, L.S., Connell, M.E., Abduljaleel, N.: Hindi CLIR in thirty days. *ACM Transactions on Asian Language Information Processing (TALIP)* 2(2), 130–142 (2003)
15. Pingali, P., Jagarlamudi, J., Varma, V.: Experiments in Cross Language Query Focused Multi-Document Summarization. In: *IJCAI 2007 Workshop on CLIA*, Hyderabad, India (2007)
16. Pingali, P., Varma, V., Tune, K.K.: Evaluation of Oromo-English Cross-Language Information Retrieval. In: *IJCAI 2007 Workshop on CLIA*, Hyderabad, India (2007)
17. Philips, L.: The Double-Metaphone Search Algorithm. *C/C++ User's Journal* 18(6) (2000)
18. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281. ACM Press, New York (1998)
19. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. In: *Behavior Research Methods, Instrumentation, and Computers*, pp. 203–208 (1996)

Amharic-English Information Retrieval

Atelach Alemu Argaw and Lars Asker

Department of Computer and Systems Sciences, Stockholm University/KTH
{[atelach,asker](mailto:atelach,asker@dsv.su.se)}@dsv.su.se

Abstract. We describe Amharic-English cross lingual information retrieval experiments in the ad hoc bilingual tracks of the CLEF 2006. The query analysis is supported by morphological analysis and part of speech tagging while we used two machine readable dictionaries supplemented by online dictionaries for term lookup in the translation process. Out of dictionary terms were handled using fuzzy matching and Lucene[4] was used for indexing and searching. Four experiments that differed in terms of utilized fields in the topic set, fuzzy matching, and term weighting, were conducted. The results obtained are reported and discussed.

1 Introduction

Amharic is the official government language spoken in Ethiopia. It is a Semitic language of the Afro-Asiatic Language Group that is related to Hebrew, Arabic, and Syrian. Amharic, a syllabic language, uses a script which originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). The language has 33 basic characters with each having 7 forms for each consonant-vowel combination, and extra characters that are consonant-vowel-vowel combinations for some of the basic consonants and vowels. It also has a unique set of punctuation marks and digits. Unlike Arabic, Hebrew or Syrian, the language is written from left to right. Amharic alphabets are one of a kind and unique to Ethiopia.

Manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literature, journalism, education, national business and cross-communication. A wide variety of literature including religious writings, fiction, poetry, plays, and magazines are available in the language.

The Amharic topic set for CLEF 2006 was constructed by manually translating the English topics. This was done by professional translators in Addis Ababa. The Amharic topic set which was written using 'fidel', the writing system for Amharic, was then transliterated to an ASCII representation using SERA¹. The transliteration was done using a file conversion utility called g2² which is available in the LibEth³ package.

¹ SERA stands for System for Ethiopic Representation in ASCII,
<http://www.abysyniacybergateway.net/fidel/sera-faq.html>

² g2 was made available to us through Daniel Yacob of the Ge'ez Frontier Foundation
(<http://www.ethiopic.org/>)

³ LibEth is a library for Ethiopic text processing written in ANSI C <http://libeth.sourceforge.net/>

We designed four experiments in our task. The experiments differ from one another in terms of fuzzy matching, term weighting, and usage of the title and description fields in the topic sets. Details of these are given in Section 4. Lucene [4], an open source search toolbox, was used as the search engine for these experiments.

The paper is organized as follows, Section 1 gives an introduction of the language under consideration and the overall experimental setup. Section 2 deals with the query analysis which consists of morphological analysis, part of speech tagging, filtering as well as dictionary lookup. Section 3 reports how out of dictionary terms were handled. It is followed by the setup of the four retrieval experiments in section 4. Section 5 presents the results and section 6 discusses the obtained results and gives concluding remarks.

2 Query Analysis and Dictionary Lookup

Dictionary lookup requires that the (transliterated) Amharic terms are first morphologically analyzed and represented by their lemmatized citation form. Amharic, just like other Semitic languages, has a very rich morphology. A verb could for example have well over 150 different forms. This means that successful translation of the query terms using a machine readable dictionary will be crucially dependent on a correct morphological analysis of the Amharic terms.

For our experiments, we developed a morphological analyzer and Part of Speech (POS) tagger for Amharic. The morphological analyzer finds all possible segmentations of a given word according to the morphological rules of the language and then selects the most likely prefix and suffix for the word based on corpus statistics. It strips off the prefix and suffix and then tries to look up the remaining stem (or alternatively, some morphologically motivated variants of it) in a dictionary to verify that it is a possible segmentation. The frequency and distribution of prefixes and suffixes over Amharic words is based on a statistical analysis of a 3.5 million word Amharic news corpus. The POS-tagger, which is trained on a 210,000 words, manually tagged Amharic news corpus, selects the most likely POS-tag for unknown words, based on their prefix and suffix combination as well as on the POS-tags of the other words in the sentence.

The output from the morphological analyzer is used as the first pre-processing step in the retrieval process. We used the morphological analyzer to lemmatize the Amharic terms and the POS-tagger to filter out less content bearing words. The morphological analyzer had an accuracy of around 86% and the POS tagger had an accuracy of approximately 97% on the 50 queries in the Amharic topic set. After the terms in the queries were POS tagged, the filtering was done by keeping Nouns and Numbers in the keyword list being constructed while discarding all words with other POS tags.

Starting with tri-grams, bi-grams and finally at the word level, each term was then looked up in the an Amharic - English dictionary [2]. For the tri- and bi-grams, the morphological analyzer only removed the prefix of the first word and the suffix of the last word respectively. If the term could not be found in

the dictionary, a triangulation method is used where by the terms were looked up in an Amharic - French dictionary [1] and sub sequentially the terms are translated from French to English using an on-line English - French dictionary WordReference (<http://www.wordreference.com/>). We also used an on-line English - Amharic dictionary (<http://www.amharicdictionary.com/>) to translate the remaining terms that were not found in any of the above dictionaries.

For the terms that were found in the dictionaries, we used all senses and all synonyms that were found. This means that one single Amharic term could in our case give rise to as many as up to eight alternative or complementary English terms. At the query level, this means that each query was initially maximally expanded.

3 Out-of-Dictionary Terms

Those terms that where POS-tagged as nouns and not found in any of the dictionaries were selected as candidates for possible fuzzy matching using edit distance. The assumption here is that these words are most likely cognates, named entities, or borrowed words. The candidates were first filtered by their frequency of occurrence in a large (3.5 million words) Amharic news corpus. If their frequency in the corpus (in either their lemmatized or original form) is more than a predefined threshold value of 10^4 , they would be considered likely to be non-cognates, and removed from the fuzzy matching unless they were labeled as cognates by a classifier specifically trained to find (English) cognates and loan words in Amharic text. The classifier is trained on 500 Amharic news texts where each word has been manually labeled as to whether it is a cognate or not. The classifier, which has a precision of around 98% for identifying cognates, is described in more detail in [3].

The set of possible fuzzy matching terms was further reduced by removing those terms that occurred in 9 or more of the original 50 queries assuming that they would be remains of non informative sentence fragments of the type "Find documents that describe...". When the list of fuzzy matching candidates had been finally decided, some of the terms in the list were slightly modified in order to allow for a more "English like" spelling than the one provided by the transliteration system [5]. All occurrences of "x" which is a representation of the sound 'sh' in the Amharic transliteration, would be replaced by "sh" ("jorj bux" → "jorj bush").

4 Retrieval

The retrieval was done using the Apache Lucene, an open source high-performance full-featured text search engine library written in Java [4]. It is a technology

⁴ It should be noted that this number is an empirically set number and is dependent on the type and size of the corpus under consideration.

deemed suitable for applications that require full-text search, especially in a cross-platform.

There are clearly a number of factors that can influence the performance in a cross lingual retrieval task. The performance of the search engine, the accuracy of the morphological analyzer, the quality and completeness of the lexicon, the effectiveness of the query expansion and the word sense disambiguation steps, or the handling of out-of-dictionary terms are just a few of the factors that will influence the overall system performance. The large number of parameters to be varied and their possible combinations makes overall system optimization practically impossible. Instead, the search for a reasonably well tuned retrieval system is directed by availability of resources, heuristic knowledge (nouns tend to carry more information than e.g. prepositions or conjunctions), and in some cases univariate sensitivity tests aimed at optimizing a specific single parameter while keeping the others fixed at reasonable values.

The four runs that were submitted this year were produced as a result of experiments aimed at investigating the effect of a few of these factors: the importance of out-of-dictionary terms, the relative importance of query length in terms of using the fields Title and Description, and the effectiveness of the fuzzy matching step. In general, for all four runs, we used only terms that had been labeled as nouns or numbers by the POS-tagger. We also give each term a default weight of 1 unless otherwise specified in the description of the individual run. Further details of the four runs are described in more detail below.

4.1 First Run R1_full

This run used the fully expanded query terms from both the Title and Description fields of the Amharic topic set. In order to cater for the varying number of synonyms that are given as possible translations for each of the query terms, the corresponding translated term sets for each Amharic term were down weighted. This was done by dividing 1 by the number of terms in each translated set and giving those equal fractional weights that adds up to 1. An Amharic term that had only one possible translation according to the dictionary would thus assign a weight of 1 to the translated term while an Amharic term with four possible translations would assign a weight of 0.25 to each translated term. All possible translations are given equal weight and there is no discrimination for synonyms and translational polysemes. The importance of the down weighting lies in the assumption that it will help control the weight each term in the original query contributes. It is also assumed that such down weighting will have a positive effect against possible query drift. An edit distance based fuzzy matching was used in this experiment to handle cognates, named entities and borrowed words that were not found in any of the dictionaries. The edit distance fuzzy matching used an empirically determined threshold value of 0.6 to restrict the matching.

4.2 Second Run R2_title

In the second run the conditions are similar to the ones in the first run except that only the terms from the Title field of the Amharic topic set were used.

The underlying assumption here is that the title of a document (which can be seen as an extreme summary of the full text) is more likely to contain a higher concentration of content bearing words. This is an attempt to investigate how much the performance of the retrieval is affected with varying query length.

4.3 Third Run R3_weight

In the third run, terms from both the Title and Description fields were used. The difference from the first run is that terms that were not found in any of the dictionaries were given a much higher importance in the query set by boosting their weight to 10. The reason for this is that words that are not found in the dictionary are much more likely to be proper names or other words that are assumed to carry a lot of information. The boosting of weights for these terms is intended to highlight the assumed strong effect that these terms will have on retrieval performance. It is of course very important that the terms selected for fuzzy matching contain as few occurrences as possible of terms that have incorrectly been assumed to be cognates, proper names or loan words. With the parameter settings that we used in our experiments, only one out of the 60 terms that were selected for fuzzy matching was a non-cognate.

4.4 Fourth Run R4_nofuzz

The fourth run is designed to be used as a comparative measure of how much the fuzzy matching affects the performance of the retrieval system. The setup in the first experiment is adopted here, except for the use of fuzzy matching. Those words that were not found in the dictionary, which so far have been handled by fuzzy matching, were manually translated into their corresponding English term in this run.

5 Results

Table 1 lists the precision at various levels of recall for the four runs. As can be seen from both Table 1 and Table 2, the best performance is achieved for run R4_nofuzz where the filtered out-of-dictionary words have been manually translated into their respective appropriate English term. We see from Table 1 that the run with the manually translated terms has a performance that is approximately 4 - 6 % better for most levels of recall when compared to the best performing automatic run, R1_full.

When comparing the results for the first two runs (R1_full and R2_title) there is a significant difference in performance. The title fields of the topic set are mainly composed of named entities that would be unlikely to be found in machine readable dictionaries. In our approach, this implies that the search using the title fields only relies heavily on fuzzy matching, which in the worst cases, would match to completely unrelated words. When using both the title

and the description fields, the named entities are supplemented by phrases that describe them. These phrases could serve as the 'anchors' that keep the query from drifting due to incorrect fuzzy matching. We performed a query by query analysis of the results given by short queries (title) vs longer queries (title + description). In general, queries which have a better or approximately equal mean average precision in relation to the median for long queries perform similarly when using the corresponding short queries while they tend to have a more negative effect for queries that have a much lower average precision than the median.

The third run, **R3_weight**, that uses a boosted weight for those terms that are not found in the dictionary, performed slightly worse than the first run, **R1_full**, on average the performance is around 2% lower than the first run. Boosting the weight of named entities in queries is expected to give better retrieval performance since such words are expected to carry a great deal of information. But in setups like ours where these named entities rely on fuzzy matching the boosting has a positive effect for correct matches as well as a highly negative effect for incorrect matches.

Table 1. Recall-Precision tables for the four runs

Recall	R1_full	R2_title	R3_weight	R4_nofuzz
0.00	40,90	31,24	38,50	47,19
0.10	33,10	25,46	28,35	39,26
0.20	27,55	21,44	23,73	31,85
0.30	24,80	18,87	21,01	28,61
0.40	20,85	16,92	16,85	25,19
0.50	17,98	15,06	15,40	23,47
0.60	15,18	13,25	13,24	20,60
0.70	13,05	11,73	10,77	17,28
0.80	10,86	8,49	8,50	14,71
0.90	8,93	6,85	6,90	11,61
1.00	7,23	5,73	6,05	8,27

A summary of the results obtained from all runs is reported in Table 2. The number of relevant documents, the retrieved relevant documents, the non-interpolated average precision as well as the precision after R (where R is the

Table 2. Summary of results for the four runs

	<i>Relevant-tot</i>	<i>Relevant-retrieved</i>	<i>Avg Precision</i>	<i>R-Precision</i>
R1_full	1,258	751	18.43	19.17
R2_title	1,258	643	14.40	16.47
R3_weight	1,258	685	15.70	16.60
R4_nofuzz	1,258	835	22.78	22.83

number of relevant documents for each query) documents retrieved (R-Precision) are summarized in the table.

6 Discussion and Directives

We have been able to get better retrieval performance for Amharic compared to automatic runs in the previous two years. Linguistically motivated approaches were added in the query analysis. The topic set has been morphologically analyzed and POS tagged. Both the analyzer and POS tagger performed reasonably well when used to analyze the Amharic topic set. It should be noted that these tools have not been tested for other domains. The POS tags were used to remove non-content bearing words while we used the morphological analyzer to derive the citation forms of words. The morphological analysis ensured that various forms of a word would be properly reduced to the citation form and be looked up in the dictionary rather than being missed out and labeled as an out-of-dictionary entry. Although that is the case, in the few times the analyzer segments a word wrongly, the results are very bad since it entails that the translation of a completely unrelated word would be in the keywords list. Especially for shorter queries, this could have a great effect. For example in query C346, the phrase *grand slam*, the named entity *slam* was analyzed as *s-lam*, and during the dictionary look up *cow* was put in the keywords list since that is the translation given for the Amharic word *lam*. We had a below median performance on such queries.

On the other hand, stop word removal based on POS tags by keeping the nouns and numbers only worked well. Manual investigation showed that the words removed are mainly non-content bearing words. The experiment with no fuzzy matching gave the highest result since all cognates, named entities and borrowed words were added manually. From the experiments that were done automatically, the best results obtained is for the experiment with the fully expanded queries with down weighting and using both the title and description fields, while the worst one is for the experiment in which only the title fields were used. The experiment where fuzzy matching words were boosted 10 times gave slightly worse results than the non-boosted experiment. The assumption here was that such words that are mostly names and borrowed words tend to contain much more information than the rest of the words in the query. Although this may be intuitively appealing, there is room for boosting the wrong words. In such huge data collections, it is likely that there would be unrelated words matching fuzzily with those named entities. The decrease in performance in this experiment when compared to the one without fuzzy match boosting could be due to up weighting such words. Further experiments with different weighting schemes, as well as different levels of natural language processing will be conducted in order to investigate the effects such factors have on the retrieval performance.

References

1. Abebe, B.: Dictionnaire Amharique-Francais
2. Aklilu, A.: Amharic English Dictionary
3. Hagman, J.: Mining for cognates. Master's Thesis, Department of Computer and Systems Sciences, Stockholm University (2006)
4. (2005), <http://lucene.apache.org/java/docs/index.html>
5. Yacob, D.: System for ethiopic representation in ascii (sera) (1996), <http://www.abyssiniacybergateway.net/fidel/>

The University of Lisbon at CLEF 2006 Ad-Hoc Task

Nuno Cardoso, Mário J. Silva, and Bruno Martins

Faculty of Sciences, University of Lisbon
{ncardoso,mjs,bmartins}@xldb.di.fc.ul.pt

Abstract. This paper reports the participation of the XLDB Group from the University of Lisbon in the CLEF 2006 ad-hoc monolingual and bilingual subtasks for Portuguese. We present our IR system, detail the query expansion strategy and the weighting scheme, describe the submitted runs and discuss the obtained results.

1 Introduction

This paper describes the third participation of the XLDB Group from the University of Lisbon in the CLEF ad-hoc task. Our main goal was to obtain a stable platform to test GIR approaches for GeoCLEF task [1].

In 2004 we participated with an IR system made from components of *tumba!*, our web search engine [2]. We learnt that searching and indexing large web collections is different than querying CLEF ad-hoc newswire collections [3]. Braschler and Peters overviewed the best IR systems of the CLEF 2002 campaign and concluded that they relied on robust stemming, a good term weighting scheme and a query expansion approach [4]. *Tumba!* does not have a stemming module and does not perform query expansion. Its weighting scheme, built for web documents, is based on PageRank [5] and in HTML markup elements. As a result, we needed to develop new modules to properly handle the ad-hoc task.

In 2005 we developed *QuerCol*, a query generator module with query expansion, and implemented a $tf \times idf$ term weighting scheme with a result set merging module for our IR system [6]. The results improved, but were still far from our performance goal. This year we improved *QuerCol* with a blind relevance feedback algorithm, and implemented a term weighting scheme based on BM25 [7].

2 IR System Architecture

Figure 1 presents our IR system architecture. In the data loading step, the CLEF collection is loaded in a repository, so that *SIDRA*, the indexing system of *tumba!*, can index the collection and generate term indexes. For our automatic runs, *QuerCol* loads the CLEF topics and generates query strings. In the retrieval step, the queries are submitted to *SIDRA* through a run generator, producing runs in CLEF format. In the rest of this Section we detail the modules shadowed in grey, *QuerCol* and *SIDRA*.

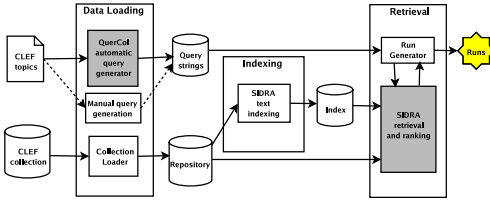


Fig. 1. The IR system architecture

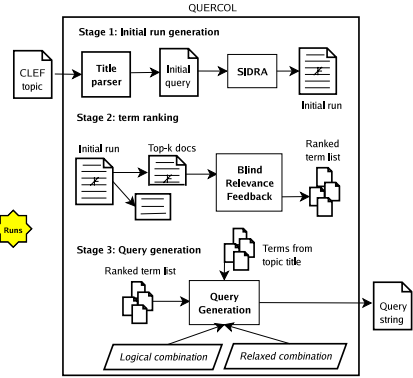


Fig. 2. Details of the QuerCol module

2.1 QuerCol Query Generator

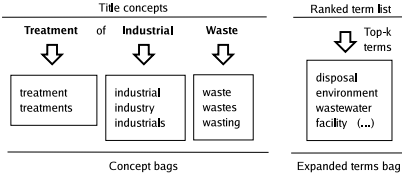
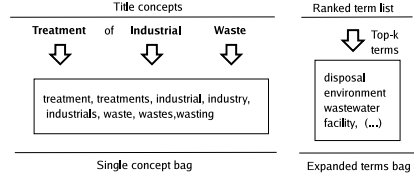
This year, we improved QuerCol with a query expansion step using blind relevance feedback [8,9]. Together with a query construction step, QuerCol can parse CLEF topics and generate query strings without human intervention. QuerCol operates in three stages (see Figure 2):

Stage 1: Initial run generation. For each topic, the non-stopword terms from its title are extracted and combined as Boolean AND expressions, yielding the queries submitted to SIDRA that generated the initial runs. Note that, in our automatic runs, we did not use the description or the narrative fields.

Stage 2: Term ranking. We used the $w_t(p_t - q_t)$ algorithm to weight the terms for our query expansion algorithm [10]. QuerCol assumes that only the documents above a certain threshold parameter, the *top-k documents*, are relevant for a given topic. The top-k documents are then tokenised and their terms are weighted, generating a *ranked term list* that best represents the top-k documents.

Stage 3: Query generation. QuerCol combines terms from the ranked term list and from the topic title, to generate a Boolean query string in the Disjunctive Normal Format (DNF). This combination may use the AND expression (operation $\times_{AND}$) or the OR expression (operation $\times_{OR}$). We defined two query construction approaches, the *Logical Combination* (LC) and the *Relaxed Combination* (RC).

The *Logical Combination* assumes that all non-stopwords from the topic title are different *concepts* that must be mentioned in the retrieved documents (see Figure 3). The generated query string forces each query to contain at least one term from each concept, reproducing the Boolean constraints described in [11].

**Fig. 3.** Logical Combination (LC) approach**Fig. 4.** Relaxed Combination (RC) approach

As each concept may be represented by many terms, the LC approach searches the ranked term list to find related terms for each concept. When found, the terms are moved into the corresponding concept's bag of terms, the *concept bags*. QuerCol relates the term to a concept if they share the same stem, as given by Porter's stemming algorithm for Portuguese [12]. After filling all concept bags, the remaining top-k terms from the ranked term list are moved into a new bag, called *expanded terms bag*. This bag contains terms that are strongly related to the topic, but do not share the same stem from any of the concepts.

The next stage generates all possible combinations of the $m \times n$ matrix of m bags $\times n$ terms in each bag using the $\times_{AND}$ operation, producing $m \times n$ *partial queries* containing one term from each content bag and from the expanded terms bags. The partial queries are then combined with the $\times_{OR}$ operation, resulting in a query string ($\times_{OR}[partial\ queries_{concept + expanded\ terms\ bags}]$). We found that these query strings were vulnerable to query drift, so we generated an additional query string using only the concept bags in a similar way ($\times_{OR}[partial\ queries_{concept\ bags}]$), and then combined both query strings using an $\times_{OR}$ operation. In DNF, the final query string generated by the LC approach is the following:

$$\times_{OR} [\times_{OR}[partial\ queries_{concept\ bags}], \times_{OR}[partial\ queries_{concept + expanded\ terms\ bags}]]$$

Nonetheless, there are relevant documents that may not mention all concepts from the topic title [11]. As the LC forces all concepts to appear in the query strings, we may not retrieve some relevant documents. Indeed, last year QuerCol generated query strings following an LC approach, and we observed low recall values in our results [6].

To tackle the limitations of the LC, we implemented a modified version, called the *Relaxed Combination*. The RC differs from the LC by using a single bag to collect the related terms (the *single concept bag*), instead of a group of concept bags (see Figure 4). This modification relaxes the Boolean constraints from the LC. The RC approach generates partial queries in a similar way as the LC approach, combining terms from the two bags (the single concept bag and the expanded terms bag) using the $\times_{AND}$ operation. The partial queries are then combined using the $\times_{OR}$ operation, to generate the final query string. In DNF, the RC approach is the following:

$$\times_{OR} [partial\ queries_{single\ concept + expanded\ terms\ bags}]$$

2.2 Weighting and Ranking

The SIDRA retrieval and ranking module implemented the BM25 weighting scheme. The parameters were set to the standard values of $k_1=2.0$ and $k_2=0.75$. Robertson et al. proposed an extension of the BM25 scheme for structured documents, suggesting that document elements such as the title could be repeated in a corresponding unstructured document, so that the title terms can be weighted more important [13].

For CLEF, we assumed that the first three sentences of each document should be weighted as more important, as the first sentences of news articles often contains a summary of the content. Robertson’s extension was applied to generate run PT4, giving a weight of 3 to the first sentence, and a weight of 2 to the following two sentences. SIDRA’s ranking module was also improved to support disjunctive queries more efficiently, so we abandoned the result sets merging module that we developed for CLEF 2005 [6].

3 Results

We submitted four runs for the Portuguese ad-hoc monolingual subtask and four other runs for the English to Portuguese ad-hoc bilingual subtask. The Portuguese monolingual runs evaluated both the QuerCol query construction strategies and the BM25 term weight extension, while the English runs evaluated different values for the top ranked documents threshold (top-k documents), and for the size of the expanded terms bag (top-k terms). Table 1 summarises the configuration of the submitted runs.

Run PT1 was manually created from topic terms, their synonyms and morphological expansions. For our automatic runs, we used the CLEF 2005 topics and qrels to find the best top-k term values for a fixed value of 20 top-k documents. The LC runs obtained a maximum MAP value of 0.2099 for 8 top-k terms. The RC runs did not perform well for low top-k term values, but for higher values they outperformed the LC runs with a maximum MAP value of 0.2520 for 32 top-k terms. For the English to Portuguese bilingual subtask, we translated the topics with Babelfish (<http://babelfish.altavista.com>) and tested with half of the top-k terms and top-k documents, to evaluate if they significantly affect the results.

Table 2 presents our results. For the Portuguese monolingual subtask, we observe that our best result was obtained by the manual run, but the automatic runs achieved a performance comparable to the manual run. The RC produced better results than the LC, generating our best automatic runs. The BM25 extension implemented in the PT4 run did not produce significant improvements.

Table 1. Runs submitted for the Portuguese (PT) and English (EN) monolingual

Label	Type	Query construction	Top-k terms	Top-k docs	BM25 extension	Label	Type	Query construction	Top-k terms	Top-k docs	BM25 extension
PT1	Manual	Manual	-	20	no	EN1	Automatic	Relaxed	16	10	no
PT2	Automatic	Logical	8	20	no	EN2	Automatic	Relaxed	32	10	no
PT3	Automatic	Relaxed	32	20	no	EN3	Automatic	Relaxed	16	20	no
PT4	Automatic	Relaxed	32	20	yes	EN4	Automatic	Relaxed	32	20	no

Table 2. Overall results for all submitted runs

Measure	PT1	PT2	PT3	PT4	EN1	EN2	EN3	EN4
num_q	50	50	50	50	50	50	50	50
num_ret	13180	7178	48991	49000	41952	42401	42790	43409
num_rel	2677	2677	2677	2677	2677	2677	2677	2677
num_rel_ret	1834	1317	2247	2255	1236	1254	1275	1303
map	0,3644	0,2939	0,3464	0,3471	0,2318	0,2371	0,2383	0,2353
gm_ap	0,1848	0,0758	0,1969	0,1952	0,0245	0,0300	0,0377	0,0364
R-prec	0,4163	0,3320	0,3489	0,3464	0,2402	0,2475	0,2509	0,2432
bpref	0,3963	0,3207	0,3864	0,3878	0,2357	0,2439	0,2434	0,2362
recip_rank	0,7367	0,7406	0,6383	0,6701	0,4739	0,4782	0,5112	0,4817

For the English to Portuguese bilingual task, we observe that the different top-k terms and top-k document values do not affect significantly the performance of our IR system. The PT3 and EN4 runs were generated with the same configuration, to compare our performance in both subtasks. The monolingual run obtained the best result, with a difference of 32% in the MAP value to the corresponding bilingual run.

4 Conclusion

This year, we implemented well-known algorithms in our IR system to obtain good results on the ad-hoc task, allowing us to stay focused on GIR approaches for the Geo-CLEF task. Our results show that we improved our monolingual IR performance in both precision and recall. The best run was generated from a query built with a Relaxed Combination, with an overall recall value of 84.2%. We can not tell at this time what are the contributions of each module to the achieved improvements of the results.

The English to Portuguese results show that the topic translation was poor, resulting in a decrease of 0.111 in the MAP values for runs PT3 and EN4. The difference between the two runs shows that we need to adopt another strategy to improve our bilingual results. We also observe that the top-k terms and top-k document values did not affect significantly the performance of the IR system.

Next year, our efforts should focus on improving the query expansion and query construction algorithms. QuerCol can profit from the usage of the description and narrative fields, producing better query strings. Also, we can follow the suggestion of Mitra et al. and rerank the documents before the relevance feedback, to ensure that the relevant documents are included in the top-k document set [11].

Acknowledgements. We would like to thank Daniel Gomes who built the tumba! repository, Leonardo Andrade for developing SIDRA, Alberto Simões for topic translations, and to all developers of tumba!. Our participation was partially supported by grants POSI/PLP/43931/2001 (Linguatca) and POSI/SRI/40193/2001 (GREASE) from FCT (Portugal), co-financed by POSI.

References

1. Martins, B., Cardoso, N., Chaves, M., Andrade, L., Silva, M.J.: The University of Lisbon at GeoCLEF 2006. In: Peters, C., ed.: Working Notes for the CLEF 2006 Workshop, Alicante, Spain (2006)

2. Silva, M.J.: The Case for a Portuguese Web Search Engine. In: Proceedings of ICWI-03, the 2003 IADIS International Conference on WWW Internet, Algarve, Portugal, pp. 411–418.
3. Cardoso, N., Silva, M.J., Costa, M.: The XLDB Group at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 245–252. Springer, Heidelberg (2005)
4. Braschler, M., Peters, C.: Cross-language evaluation forum: Objectives, results, achievements. *Information Retrieval* 7, 7–31 (2004)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library (1999)
6. Cardoso, N., Andrade, L., Simões, A., Silva, M.J.: The XLDB Group participation at CLEF 2005 ad hoc task. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 54–60. Springer, Heidelberg (2006)
7. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M.: Okapi at TREC-3. In IST Special Publication 500-225. In: Harman, D. (ed.): Overview of the Third Text REtrieval Conference (TREC 3), Gaithersburg, MD, USA, Department of Commerce, National Institute of Standards and Technology, pp. 109–126 (1995)
8. Rocchio Jr., J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice-Hall, Englewood Cliffs (1971)
9. Efthimiadis, E.N.: Query Expansion. 31, 121–187 (1996)
10. Efthimiadis, E.N.: A user-centered evaluation of ranking algorithms for interactive query expansion. In: Proceedings of ACM SIGIR '93, pp. 146–159. ACM Press, New York (1993)
11. Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 206–214. ACM Press, New York (1998)
12. Porter, M.: An algorithm for suffix stripping. *Program* 14, 130–137 (1980)
13. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: CIKM '04: Proceedings of the thirteenth ACM international Conference on Information and Knowledge Management, pp. 42–49. ACM Press, New York (2004)

Query and Document Translation for English-Indonesian Cross Language IR

Herika Hayurani, Syandra Sari, and Mirna Adriani

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
{heha51, sysa51}@ui.edu, mirna@cs.ui.ac.id

Abstract. We present a report on our participation in the Indonesian-English ad hoc bilingual task of the 2006 Cross-Language Evaluation Forum (CLEF). This year we compare the use of several language resources to translate Indonesian queries into English. We used several readable machine dictionaries to perform the translation. We also used two machine translation techniques to translate the Indonesian queries. In addition to translating an Indonesian query set into English, we also translated English documents into Indonesian using the machine readable dictionaries and a commercial machine translation tool. The results show performing the task by translating the queries is better than translating the documents. Combining several dictionaries produced better result than only using one dictionary. However, the query expansion that we applied to the translated queries using the dictionaries reduced the retrieval effectiveness of the queries.

1 Introduction

This year we participate in the bilingual 2006 Cross Language Evaluation Forum (CLEF) ad hoc task, i.e., the English-Indonesian CLIR. As stated in previous work [8], a translation step must be done either to the documents [9] or to the queries [3, 5, 6] in order to overcome the language barrier. The translation can be done using bilingual readable machine dictionaries [1, 2], machine translation techniques [7], or parallel corpora [11]. We used a commercial machine translation software package called *Transtool*¹ and an online machine translation system available on the Internet to translate an Indonesian query set into English and to translate English documents into Indonesian. We learned from our previous work [1, 2] that freely available dictionaries on the Internet could not correctly translate many Indonesian terms, as their vocabulary was very limited. We hoped that using machine translation techniques and parallel documents could improve our result this time.

2 The Translation Process

In our participation, we translated English queries and documents using dictionaries and machine translation techniques. We manually translated the original CLEF query

¹ See “<http://www.geocities.com/cdpenerjemah/>”

set from English into Indonesian. We then translated the resulting Indonesian queries back into English using dictionaries and machine translation techniques. The dictionaries came from several sources, particularly, the *Indonesian National Research and Technology Office* (*Badan Pengkajian dan Penelitian Teknologi* or BPPT) and online dictionaries found on the internet, namely <http://www.orisinil.com/kamus.php> and <http://www.kamus.net/main.php>. The machine translation systems that we used were *Toggetext* (www.toggetext.com) and *Transtool*, a commercial software package.

Besides translating the queries, we also translated the English documents from CLEF into Indonesian. The translation process was done using a combined dictionary (the online dictionaries and the BPPT dictionary) and a machine translation (*Transtool*). The dictionary-based translation process was done by taking only the first definition for each English word in the document. It took four months to translate all the English documents in the collection into Indonesian.

2.1 Query Expansion Technique

Adding translated queries with relevant terms, known as query expansion, has been shown to improve CLIR effectiveness [1, 3, 12]. One of the query expansion techniques is called the *pseudo relevance feedback* [4, 5]. This technique is based on an assumption that the top few documents initially retrieved are indeed relevant to the query, and so they must contain other terms that are also relevant to the query. The query expansion technique adds such terms into the previous query. We apply this technique to the queries in this work. To choose the relevant terms from the top ranked documents we employ the *tf*idf* term weighting formula [10]. We added a certain number of terms that have the highest weight scores.

3 Experiment

We participated in the bilingual task with English topics. The English document collection contains 190,604 documents from two English newspapers, the *Glasgow Herald* and the *Los Angeles Times*. We opted to use the query title and the query description provided with the query topics. The query translation process was performed fully automatic using a machine translation (*Transtool*) and several dictionaries.

We then applied a pseudo relevance-feedback query-expansion technique to the queries that were translated using the machine translation tool and dictionaries. We used the top 10 relevant documents retrieved for a query from the collection to extract the expansion terms. The terms were then added to the original query. In these experiments, we used *Lemur*² information retrieval system which is based on the *language model* to index and retrieve the documents.

4 Results

Our work focused on the bilingual task using Indonesian queries to retrieve documents in the English collections. The results of our CLIR experiments were obtained

² See “<http://www.lemurproject.org/>”

Table 1. Mean average precision in CLIR runs of Indonesian queries translated into English using the machine translation tool for title only and the combination of title and description

Query / Task	Title (MAP)	Title+Description (MAP)
Monolingual	0.2824	0.3249
Translated by MT-1(Toggletext)	0.2538 (-10.12%)	0.2934 (-10.86%)
Translated by MT-2(Transtool)	0.2137 (-24.32%)	0.2397 (-26.22%)

by applying the three methods of translation, i.e., using the machine translations, using dictionaries, and using parallel corpus.

Table 1 shows the result of the first technique. It shows that translating the Indonesian queries into English using *Toggletext* machine translation tool is better than using *Transtool*. The retrieval performance of the title-based queries translated using *Toggletext* dropped 10.12% and using *Transtool* dropped 24.32% below that of the equivalent monolingual retrieval (see Table 1). The retrieval performance of the combination of query title and description translated using *Toggletext* dropped 10.86% and using *Transtool* dropped 26.22% below that of the equivalent monolingual queries.

The result of the second technique, which is translating the Indonesian queries into English using dictionaries, is shown in Table 2. The result shows that using the dictionary from BPPT alone decreased the retrieval performance of the title-based translation queries by 49.61%, and using the combined dictionaries it dropped by 26.94% below that of the equivalent monolingual retrieval. For the combined title and description queries, the retrieval performance of the translated queries using one dictionary only dropped 61.01% and using the combined dictionaries dropped 18.43% below that of the equivalent monolingual retrieval.

The last technique that we used is retrieving documents from parallel corpus created by translating the English document collection into Indonesian using the combined dictionaries and the machine translation (*Transtool*). The results, as shown in

Table 2. Mean average precision in CLIR runs of Indonesian queries translated into English using dictionaries for title only and the combination of title and description

Query / Task	Title (MAP)	Title+Description (MAP)
Monolingual	0.2824	0.3249
Dic-1	0.1423 (-49.61%)	0.1101 (-61.01%)
Dic-2 (combined)	0.2063 (-26.94%)	0.2650 (-18.43%)

Table 3. Mean average precision in the monolingual runs for title only and the combination of title and description on a parallel corpus created by translating English documents into Indonesian using dictionaries and a machine translation tool

Query / Task	Title (MAP)	Title+Description (MAP)
Monolingual	0.2824	0.3249
Parallel-DIC	0.0201 (-92.88%)	0.0271 (-91.65%)
Parallel-MT	0.2060 (-27.05%)	0.2515 (-22.59%)

Table 3, indicate that using parallel corpus created using dictionaries dropped the retrieval performance of the title only by 92.88% and the title and description queries by 91.65%. On the other hand, using the parallel corpus created by the machine translation tool dropped the retrieval performance of the title only by 27.05% and the title and description queries by 22.59%. The results indicate that retrieving the English version of Indonesian documents that are relevant to an Indonesian query is much more effective if the parallel corpus is created using the machine translation tool than using the dictionaries.

Lastly, we also attempted to improve our CLIR results by expanding the queries translated using the dictionaries. The results is as shown in Table 4, which indicates that adding the queries with 5 terms from the top-10 documents obtained from a pseudo relevance feedback technique hurt the retrieval performance of the translated queries. The query expansion that are applied to other translated queries also decreased the retrieval performance compared to the baseline results (no expansion applied).

Table 4. Mean average precision for the title only and the combination of title and description using the query expansion technique with top-10 document method

Query/Task	Title (MAP)	Title+Description (MAP)
Dic-2	0.2063	0.2650
Dic-2 + QE	0.1205 (-41.58%)	0.1829 (-30.98%)

5 Summary

The results of our experiments demonstrate that translating queries using machine translation tools is comparable than translating documents. The retrieval performance of queries that were translated using machine translation tools for Bahasa Indonesia was about 10.12%-26.22% of that of retrieving the documents translated using machine translation. Taking the first definition in the dictionary when translating an Indonesian query into English appeared to be less effective. The retrieval performance of the translated queries fall between 18.43-61.01% below that of the monolingual

queries. The result of combining several dictionaries is much better than only using one dictionary for translating the queries.

The retrieval performance of queries using a parallel corpus, created by translating the English documents into Indonesian using machine translation, is comparable to that of using the translated queries. However, retrieval using a parallel corpus that is created by translating the English documents into Indonesian using the dictionaries does not perform well. The result is 65.94-73.22% worse than that using the queries that are translated using the dictionaries.

In order to improve the retrieval performance of the translated queries, we expanded the queries with terms extracted from a number of top-documents. However, the pseudo relevance feedback technique that is known to improve the retrieval performance did not improve the retrieval performance of our queries.

References

1. Adriani, M., van Rijsbergen, C.J.: Term Similarity Based Query Expansion for Cross Language Information Retrieval. In: Abiteboul, S., Vercoustre, A.-M. (eds.) ECDL 1999. LNCS, vol. 1696, pp. 311–322. Springer, Heidelberg (1999)
2. Adriani, M.: Ambiguity Problem in Multilingual Information Retrieval. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, Springer, Heidelberg (2001)
3. Adriani, M.: English-Dutch CLIR Using Query Translation Techniques. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, Springer, Heidelberg (2002)
4. Attar, R., Fraenkel, A.S.: Local Feedback in Full-Text Retrieval Systems. *Journal of the Association for Computing Machinery* 24, 397–417 (1977)
5. Ballesteros, L., Croft, W.B.: Resolving Ambiguity for Cross-language Retrieval. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71 (1998)
6. Davis, M., Dunning, T.: A TREC Evaluation of Query Translation Methods for Multilingual Text Retrieval. In: Harman, D.K. (ed.) The Fourth Text Retrieval Conference (TREC-4), NIST (November 1995)
7. Jones, G., Lam-Adesina, A.M.: Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 105–114. Springer, Heidelberg (2002)
8. McCarley, J.S.: Should We Translate the Documents or the Queries in Cross-Language Information Retrieval? In: Proceedings of the Association for Computational Linguistics (ACL'99), pp. 208–214 (1999)
9. Oard, D.W., Hackett, P.G.: Document Translation for Cross-Language Text Retrieval at the University of Maryland. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), Gaithersburg MD (November 1997)
10. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
11. Sheridan, P., Ballerini, J.P.: Experiments in Multilingual Information Retrieval using the SPIDER System. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zürich, Switzerland (August 1996)
12. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: Proceedings in the 19th Annual International ACM SIGIR Conference on Research and development in Information Retrieval (1996)

Passage Retrieval vs. Document Retrieval in the CLEF 2006 Ad Hoc Monolingual Tasks with the IR-n System

Elisa Noguera and Fernando Llopis

Grupo de investigación en Procesamiento del Lenguaje Natural y Sistemas de
Información

Departamento de Lenguajes y Sistemas Informáticos

University of Alicante, Spain

{elisa,llopis}@dlsi.ua.es

Abstract. The paper describes our participation in monolingual tasks at CLEF 2006. We submitted results for the following languages: English, French, Portuguese and Hungarian. We focused on studying different weighting schemes (okapi and dfr) and retrieval strategies (passage retrieval and document retrieval) to improve retrieval performance. After an analysis of our experiments and of the official results at CLEF 2006, we achieved considerably improved scores by using different configurations for different languages (French, Portuguese and Hungarian).

1 Introduction

In our sixth participation at CLEF, we focused on evaluating a new weighting model (dfr), comparing retrieval strategies (based on passages/documents) and setting the best configuration to each language. Specifically, we participated in tasks for the following languages: English, French, Portuguese and Hungarian.

The IR-n system [3] was developed in 2001. It is a Passage Retrieval (PR) system which uses passages with a fixed number of sentences. This provides the passages with some syntactical content. Previous researches with the IR-n system aimed at detecting the most suitable size passage for each collection (to experiment with test collection), and at determining the similarity of a document based on the passage with most similarity. Last year, we proposed a new method, which we called *combined size passages* [4], in order to improve the performance of the system by combining different size passages. This year, we implemented a new similarity measure in the system and we tested our system with different configurations for each language.

The IR-n system uses several similarity measures. This year, the weighting model dfr [1] was included, but we also used the okapi weighting model [2]. The IR-n architecture allows us to use query expansion based on either the most relevant passages or the most relevant documents.

This paper is organized as follows: the next section describes the task developed by our system and the training carried out for CLEF 2006. The results obtained are then presented. Finally, we present the conclusions and the future work.

2 Experiments

This section describes the training process was been carried out in order to obtain the best features to improve the performance of the system. In CLEF 2006, our system participated in the following monolingual tasks: English, French, Portuguese and Hungarian.

The aim of the experimental phase was to set up the optimum value of the input parameters for each collection. CLEF-2005 (English, French, Portuguese and Hungarian) collection were used for training. Query expansion techniques were also used for all languages. Here below, we describe the input parameter of the system:

- **Size passage (sp):** We established two passage sizes: **8 sentences** (normal passage) or **30 sentences** (big passage).
- **Weighting model (wm):** We use two weighting models: **okapi** and **dfr**.
- **Opaki parameters:** these are k_1 , b and $avgld$ (k_3 is fixed as 1000).
- **Dfr parameters:** these are c and $avgld$.
- **Query expansion parameters:** If **exp** has value 1, this denotes we use relevance feedback based on passages in this experiment. But, if **exp** has value 2, the relevance feedback is based on documents. Moreover, **np** and **nd** denote the k terms extracted from the best ranked passages (np) or documents (nd) from the original query.
- **Evaluation measure:** Mean average precision (**avgP**) is the evaluation measure used in order to evaluate the experiments.

Table 1 shows the best configuration for each language. These configurations were used at CLEF 2006.

Table 1. Configurations used at CLEF 2006

language	run	sp	wm	C	avgld	k1	b	exp	np	nd	avgP
English	8-dfr-exp	8	dfr	4	600			2	10	10	0.5403
French	9-okapi-exp	9	okapi	300	1.5	0.3	1	5	10	10	0.3701
Hungarian	30-dfr-exp	30	dfr	2	300			2	10	10	0.3644
Portuguese	30-dfr-exp	30	dfr	4	300			1	5	10	0.3948

The best weighting scheme for English was dfr. We used 8 as passage size and obtained 0.5403 as average precision. For French, the best weighting scheme was okapi with 9 as passage size. Using this configuration, we obtained 0.3701 as average precision. The best weighting scheme for Hungarian was dfr, whereas the passage size was 30. The best average precision obtained with this configuration was 0.3644. Finally, the best configuration for Portuguese was the same as for Hungarian (dfr as weighting scheme and 30 as passage size). The average precision obtained with this configuration was 0.3948.

3 Results at CLEF 2006

We submitted four runs for each language in our participation (except for English for which we have submitted 1 run) in CLEF 2006. The best parameters, i.e. those that gave the best results in system training, were used in all cases.

This is the description of the runs that we submitted at CLEF 2006:

- yy-xx-zz
 - yy is the passage size
 - xx is the weighting model used (dfr or okapi)
 - zz is the expansion query (used 'exp' and not used 'nexp')

The official results for each run are showed in Table 2. Like other systems which use query expansion techniques, these models also improve performance with respect to the base system. Our results are appreciably above average in all languages, except for English where they are sensibly below average. The best percentage of improvement in AvgP is 15.43% for Portuguese.

Table 2. CLEF 2006 official results. Monolingual tasks.

Language Run		AvgP Dif	
English	CLEF Average	38.73	
	30-dfr-exp	38.17	-1.44%
French	CLEF Average	37.09	
	30-dfr-exp	37.13	
	8-dfr-exp	38.28	+3.2%
	9-okapi-exp	35.28	
	30-okapi-exp	37.80	
Portuguese	CLEF Average	37.32	
	30-dfr-exp	41.95	
	8-dfr-exp	42.06	
	8-okapi-exp	42.41	
	30-okapi-exp	43.08	+15.43%
Hungarian	CLEF Average	33.37	
	30-dfr-exp	35.32	+5.5%
	8-dfr-exp	34.25	
	30-dfr-nexp	30.60	
	8-dfr-nexp	29.50	

4 Conclusions and Future Work

In this seventh CLEF evaluation campaign, we proposed different configurations of our system for English, French, Portuguese and Hungarian (see Table 1). The best results in the training collections were obtained with these configurations. In

order to enhance retrieval performance, we evaluated different weighting models using also a query expansion approach based on passages and documents.

The results of this evaluation indicate that for the French, Portuguese and Hungarian our approach proved to be effective (see Table 2) because the results are above average. However, the results obtained for English were sensibly below average.

For Portuguese, the best results were obtained using the okapi weighting model. For other languages (English, French and Hungarian), the best results were obtained by dfr (see Table 2). Furthermore, the best passage size for French was 9, although for other languages (English, Portuguese and Hungarian) it was 30 (this passage size is comparable to IR based on the complete document).

As in previous evaluation campaigns, pseudo-relevance feedback based on passages improves mean average precision statistics for all languages, even though this improvement is not always statistically significant.

In the future we intend to test this approach on other languages such as Bulgarian and Spanish. We also intend to study ways of integrating NLP knowledge and procedures into our basic IR system and evaluating the impact.

Acknowledgements

This research has been partially supported by the framework of the project QALL-ME (FP6-IST-033860), which is a 6th Framework Research Programme of the European Union (EU), by the Spanish Government, project TEXT-MESS (TIN-2006-15265-C06-01) and by the Valencia Government under project number GV06-161.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic Models of information retrieval based on measuring the divergence from randomness. *ACM TOIS* 20(4), 357–389 (2002)
2. Savoy, J.: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) *CLEF 2003. LNCS*, vol. 3237, Springer, Heidelberg (2004)
3. Llopis, F.: IR-n: Un Sistema de Recuperación de Información Basado en Pasajes. PhD thesis, University of Alicante (2003)
4. Llopis, F., Noguera, E.: Combining Passages in the Monolingual Task with the IR-n System. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)

The PUCRS NLP-Group Participation in CLEF2006: Information Retrieval Based on Linguistic Resources

Marco Gonzalez and Vera Lúcia Strube de Lima

Grupo PLN – Faculdade de Informática – PUCRS

Av. Ipiranga, 6681 – Prédio 16 - PPGCC

90619-900 Porto Alegre, Brazil

{gonzalez, vera}@inf.pucrs.br

Abstract. This paper presents the 2006 participation of the PUCRS NLP-Group in the CLEF Monolingual Ad Hoc Task for Portuguese. We took part in this campaign using the TR+ Model, which is based on nominalization, binary lexical relations (BLR), Boolean queries, and the evidence concept. Our alternative strategy for lexical normalization, the nominalization, is to transform a word (adjective, verb, or adverb) into a semantically corresponding noun. BLRs identify relationships between nominalized terms and capture phrasal cohesion mechanisms, like those between subject and predicate, subject and object (direct or indirect), noun and adjective or verb and adverb. In our strategy, an index unit (a descriptor) may be a single term or a BLR, and we adopt the evidence concept: the descriptor weighting depends on the occurrence of phrasal cohesion mechanisms, besides depending on frequency of occurrence. We describe these features, which implement lexical normalization and term dependence in an information retrieval system based on linguistic resources.

Keywords: Search engine, information retrieval evaluation, lexical normalization, term dependence.

1 Introduction

The PUCRS NLP-Group participated in CLEF2006 in the Monolingual Ad Hoc Portuguese Retrieval Task using manual query construction from topics. Our run adopted the TR+ Model [5] based on linguistic resources.

The TR+ Model is based on nominalization [6, 7], binary lexical relations (BLRs) [4, 6], Boolean queries [5], and the evidence concept [4]. Nominalization is an alternative strategy used for lexical normalization. BLRs, which identify relationships between nominalized terms, and Boolean queries are strategies to specify term dependences. The evidence concept is part of the TR+ Model and provides a weighting scheme for the index units (henceforth called descriptors) using word frequency and phrasal cohesion mechanisms [10]. Our approach uses a probabilistic approach for information retrieval.

In our strategy, a descriptor may be a single term (e.g.: “house”) or a relationship between terms (e.g.: “house of stone”). BLRs represent those relationships (e.g.: “of(house,stone)”). A weight, an evidence value in TR+ Model, is assigned to each

descriptor (a term or BLR). The descriptor evidence value shows the importance of the respective concept that the term or the BLR describes in the text. Descriptors and their weights constitute the descriptor space.

This paper is organized as follows. Section 2 introduces the TR+ Model based on the nominalization process, the BLR recognition, the new descriptor weighting schema founded on the evidence concept, and the Boolean query formulation. Section 3 describes the difficulties found. Section 4 reports adjustments of the system, and Section 5 presents final considerations.

2 TR+ Model Overview

The TR+ Model [5] is an information retrieval model based on linguistic resources such as nominalization and term dependence recognition. Figure 1 depicts our model with its input and output, its two phases – indexing and searching phases –, and their specific steps.

In the TR+ Model documents and queries written in natural language receive the same treatment in order to construct the descriptor space (the set of index units and their weights) in the indexing phase, and to generate the input for the Boolean query formulation in the searching phase. First, in the preprocessing step we have tokenization (where words and punctuation marks are identified) and morphological tagging (where morphological tags are assigned to each word or punctuation mark). The nominalization process is then performed to generate nominalized terms and finally BLRs are extracted.

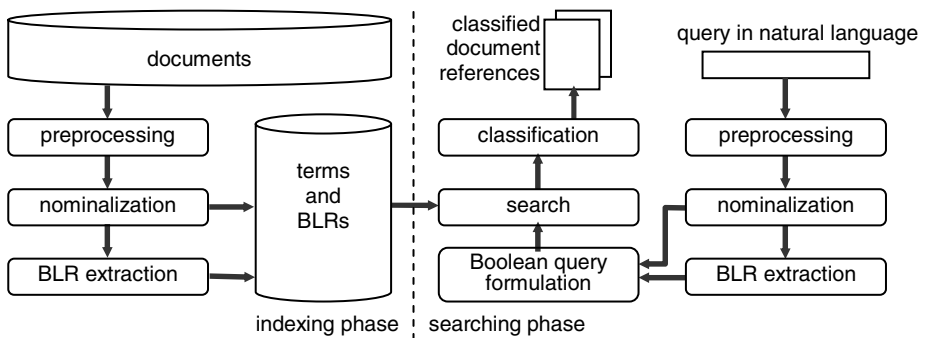


Fig. 1. TR+ Model overview

In the searching phase, we look for nominalized terms and BLRs recognized in the query, in the descriptor space. The relevance values associated with the documents for this query are computed according to descriptor weights (evidences) and to predefined Boolean operators included in the query. The final step is to classify the documents by their relevance values.

2.1 Nominalized Terms and BLRs

Morphological variation and conceptual proximity of words can have a strong impact on the effectiveness of an information retrieval system [9]. Nominalization [7] is an alternative strategy used for lexical normalization. It is based on the fact that nouns are usually the most representative words of a text [17]. Besides, queries are usually formulated through noun phrases [12]. In our work, nominalization is understood as the transformation of a word (adjective, verb, or adverb) found in the text, into a semantically corresponding noun that appears in the lexicon [8].

To put in practice this idea, an automatic nominalization process was implemented and integrated in our indexing strategy. We developed the tools FORMA [15] and CHAMA [16] that automatically derive nominalized terms from a Brazilian Portuguese text. For more details see [7, 8].

Controle controle 0 0 _SU trabalhista trabalhista trabalho 0 _AJ exato exato exatidão 0 _AJ adotado adotar adoção adotante _AP em em 0 0 _PR equipe equipe 0 0 _SU . . 0 0 _PN

Fig. 2. FORMA and CHAMA output example for Brazilian Portuguese

Figure 2 shows the output for the sentence “Controle trabalhista exato adotado em equipe” (“Exact labor control adopted in team”). The output per line in Figure 2 is organized as a list containing:

- the original word, (e.g., “adotado” (“adopted”)),
- its lemma, (e.g., “adotar” (“to adopt”)),
- the abstract noun generated (e.g., “trabalho” (“work”), “exatidão” (“accuracy”) and “adoção” (“adoption”)), when it exists, or the symbol 0, if no nominalization is produced,
- the concrete noun generated (e.g., “adotante” (“who adopts”)), when it exists, or the symbol 0, if no nominalization is produced, and
- the part-of-speech tag (in Figure 2: _SU=noun, _AJ=adjective, _AP=participle, _PR=preposition, and _PN=punctuation mark).

While nominalized terms describe atomic concepts, BLRs [4] identify relationships between them. These relationships capture phrasal cohesion mechanisms [10] from the syntactic links like those that occur between subject and predicate, subject and object (direct or indirect), noun and adjective or verb and adverb. Such mechanisms reveal term dependences.

A BLR has the form $id(t1,t2)$ where id is a relation identifier, and $t1$ and $t2$ are its arguments (nominalized terms).

We group BLRs into the following types: Classification, Restriction, and Association (prepositioned or not) [8]. We developed a tool named RELLEX [14] that automatically extracts BLRs from a Brazilian Portuguese text. For more details about

BLRs and BLR extraction rules see [4]. Those rules, and the nominalization process, are resources used to extract a unique BLR derived from different syntactic structures with the same underlying semantics.

2.2 The Evidence Concept and Descriptor Weighting

Evidence is an information that gives a strong reason for believing something or that proves something; evidences are signs, indications; something is evident if it is obvious [2, 3]. The evidence concept is crucial for the TR+ Model which adopts a descriptor weighting scheme based on this concept, i.e., the weighting is not only based on the frequency of occurrence of the descriptor. The descriptor representativeness also depends on the occurrence of phrasal cohesion mechanisms.

The evidence ($evd_{t,d}$) of a term t in a document d is:

$$evd_{t,d} = \frac{f_{t,d}}{2} + \sum_r f_{r,t,d} \quad (1)$$

where:

$f_{t,d}$ is the frequency of occurrence of t in d , and

$f_{r,t,d}$ is the number of BLRs in d where t is an argument.

On the other hand, the evidence $evd_{r,d}$ of a BLR r in a document d is:

$$evd_{r,d} = f_{r,d} (evd_{t1,d} + evd_{t2,d}) \quad (2)$$

where:

$f_{r,d}$ is the frequency of occurrence of r in d , and

$evd_{t1,d}$ and $evd_{t2,d}$ are the evidences of $t1$ and $t2$, respectively, and $t1$ and $t2$ are arguments of r .

In TR+ Model, query descriptors have their weight computed by the same formula used for documents. The evidence value substitutes the frequency of occurrence in the Okapi BM25 formula [13]. For more details see [8]. The relevance value $RV_{d,q}$ of a document d for a query q is given by:

$$RV_{d,q} = \sum_i \frac{W_{i,d} W_{i,q}}{s_i} \quad (3)$$

where:

i is a descriptor (term or a BLR);

$W_{i,d}$ is the weight for descriptor i in document d ;

$W_{i,q}$ is the weight for descriptor i in query q ; and

$s_i \neq 1$, if i is a BLR with the same arguments but different relation identifiers in d and q , or

$s_i = 1$, if i is a BLR with the same arguments and identifiers in d and q or if i is a term.

The position of each document in the resulting list depends on the relevance values, and these depend on the Boolean query formulation.

2.3 Boolean Queries and Groups of Results

A query q formulated by the user, in the TR+ Model, is treated as a text just like a document text. A Boolean query bq , automatically derived from a query q , is formulated according to the following grammar (in EBNF formalism):

```

<bq> → [ <BLRDisj> OR ] <TermConj>
<BLRDisj> → <r> [ OR <BLRDisj> ]
<TermConj> → ( <TermDisj> [ AND <TermConj> ] )
<TermDisj> → (  $\eta_1$ (<w>) OR  $\eta_2$ (<w>)) | (  $\eta_1$ (<w>)) | (  $\eta_2$ (<w>)) |
<r> → BLR
<w> → adjective | adverb | noun | verb

```

The elements η_1 and η_2 are respectively nominalization operations for generating abstract and concrete nouns. The elements OR and AND are respectively disjunction and conjunction Boolean operators. For instance, let the string “restored painting” be a query q . A corresponding Boolean query bq for this string would be:

```

“of(restoration, painting)” OR “≠of(restoration, painting)” OR ((“restoration” OR “restorer”)
AND (“painting”))

```

where “≠of(restoration, painting)” means the same arguments but different relation identifiers in d and q for the BLR (in this case $s_i = 2$ in Equation (3)).

Finally, the retrieved documents are classified in two groups:

- **Group I:** more relevant documents (that fulfill the Boolean query conditions); and
- **Group II:** less relevant documents (that do not fulfill the Boolean query conditions, but contain at least one query term).

In each of these groups, the documents are ranked in the decreasing order of their relevance value (Equation (3)).

3 Difficulties Found

This is our first participation in CLEF. Our main goal was to obtain hands-on experience in the Ad Hoc Monolingual Track on text such as the PT collections. Our prior experience was indexing and searching smaller text collections using Brazilian Portuguese only.

The changes needed in a search engine for such task are not simple adjustments and the decision to participate was taken late. We found some difficulties in the indexing phase. Our estimation is that at least 20% of the terms were not indexed due to programming errors (e.g.: mistaken variable initializations and inadequate record delimitadors).

The differences between Brazilian and European Portuguese were another source of errors because our system was designed for Brazilian Portuguese, not taking into account characteristics that are present in European Portuguese only. The following example shows this problem. While Figure 2 presents the output of our nominalization tool for a sentence in Brazilian Portuguese, Figure 3 shows the output for the same sentence (“Controlo laboral exacto adoptado em equipa”) in European Portuguese.

Controlo controlar controle controlador _VB laboral laboral laboralidade 0 _AJ exacto exacto 0 0 _SU adoptado adoptar adoptacao adoptador _AP em em 0 0 _PR equipa equipar equipamento equipador _VB . . 0 0 _PN

Fig. 3. FORMA and CHAMA output example for European Portuguese

From the example in Figure 3 it is possible to notice that the noun “Controlo” is tagged as a verb (_VB), the adjective “exacto” as a noun (_SU), and the noun “equipa” as a verb. See the Brazilian version in Figure 2. Inexistent nouns, like “laboralidade” and “adoptação”, are generated erroneously due to lexical differences between the two languages. On the other hand, nouns like “controle”, “controlador”, “equipamento”, and the wrong noun “equipador” are generated due to incorrect tagging. These mistakes affected the indexing of terms and BLRs.

4 Adjustments in the System

Some adjustments were accomplished after the CLEF2006 submission of our run trevd06. The run trevd06a (not submitted to CLEF2006) was executed after these first adjustments. They were:

- FORMA tool: recognition of different kinds of quotation marks («» and “”), special characters (e.g.: §), and accentuated characters (e.g.: ö).
- RELLEX tool: correction of programming errors found.

However, some other adjustments are still necessary. One of them is the more precise recognition and treatment of (i) some object form of pronouns (e.g.: “sugerem-no” – “they suggest him/it”), and (ii) some inflectional variations of verbs (e.g.: “aceitaríamos” – “we would accept”). These details concerning more formal writing are mandatory for the European Portuguese.

Figure 4 presents interpolated recall vs average precision for our run submitted to CLEF2006 (trevd06), and for the adjusted run (trevd06a – not submitted). Figure 4 also presents the limits of recall-precision curves for the top five participants (Top 5) in the Ad Hoc Monolingual Portuguese Track at CLEF2006 [11]. Table 1 shows some results for trevd06, trevd06a, and Top 5.

Table 1 and Figure 4 show that our model (specially concerning trevd06a) reaches good precision for the top retrieved documents (see the average precision at interpolated recall 0.0% compared to Top 5 in Table 1). The adjustments after CLEF2006 (trevd06a) improved the number of retrieved relevant documents. We hope that the future adjustments result in larger recall and MAP values.

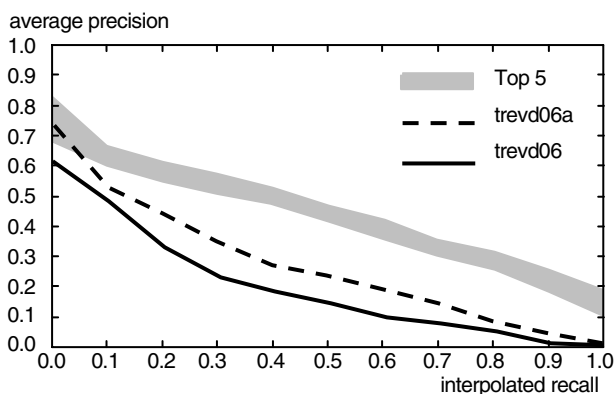


Fig. 4. Interpolated recall vs average precision

Table 1. Some results for trevd06, trevd06a, and Top 5 participants

Runs	retrieved relevant documents	not retrieved relevant documents	MAP	avg. prec. at interp. recall 0.0%
trevd06	1298	1268	18.0%	61.7%
trevd06a	1948	618	26.5%	74.0%
Top 5			40.5 – 45.5%	68.0 – 83.5%

5 Conclusion

Indeed, under the conditions of this first participation in Ad Hoc Monolingual Portuguese Track at CLEF2006, we believe that the results obtained were reasonable and the experience with European Portuguese and larger collections was extremely valid.

We have already corrected the indexing mistakes and we have observed their impact on retrieving results. There is another immediate task concerning this experience: the adaptation of our tagging and nominalization tools in order to deal with European Portuguese.

We must decide now on two work directions: (i) to use specialized tools for each Portuguese (Brazilian and European) or (ii) to create a generic tool for text preprocessing. An alternative under consideration is to transform variations of words (like “exato” and “exato” (“exact”)) into a common form, before the indexing phase.

On the other hand, we are adding other adjustments to our text preprocessing tools. They concern mainly the treatment of some object forms of pronouns and some inflectional variations of verbs. The impact due to these errors is seen in the nominalization process and in the BLR identification: they reduce the number of retrieved relevant documents.

References

1. Gamallo, P., Gonzalez, M., Agustini, A., Lopes, G., de Lima, V.L.S.: Mapping Syntactic Dependencies onto Semantic Relations. In: ECAI'02, Workshop on Natural Language Processing and Machine Learning for Ontology Engineering, Lyon, France, pp. 15–22 (2002)
2. Crowther, J.: Oxford Advanced Learner's Dictionary of Current English. Oxford University Press, New York (1995)
3. Ferreira, A.B.H.: Dicionário Aurélio Eletrônico – Século XXI. Nova Fronteira S.A., Rio de Janeiro (1999)
4. Gonzalez, M., de Lima, V.L.S., de Lima, J.V.: Binary Lexical Relations for Text Representation in Information Retrieval. In: Montoyo, A., Muñoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 21–31. Springer, Heidelberg (2005)
5. Gonzalez, M.: Termos e Relacionamentos em Evidência na Recuperação de Informação. PhD thesis, Instituto de Informática, UFRGS (2005)
6. Gonzalez, M., de Lima, V.L.S., de Lima, J.V.: Lexical normalization and term relationship alternatives for a dependency structured indexing system. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 394–405. Springer, Heidelberg (2006)
7. Gonzalez, M., de Lima, V.L.S., de Lima, J.V.: Tools for Nominalization: an Alternative for Lexical Normalization. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 100–109. Springer, Heidelberg (2006)
8. Gonzalez, M., de Lima, V.L.S., de Lima, J.V.: The PUCRS-PLN Group participation at CLEF 2006. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Working Notes for the CLEF 2006 Workshop (2006), Published Online www.clef-campaign.org
9. Krovetz, R.: Viewing morphology as an inference process. Artificial Intelligence N. 118, 227–294 (2000)
10. Mira Mateus, M.H., Brito, A.M., Duarte, I., Faria, I.H.: Gramática da Língua Portuguesa. Lisboa: Ed. Caminho (2003)
11. di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In: Nardi, A., Peters, C., Vicedo, J. L. (eds.), Working Notes for the CLEF 2006 Workshop (2006), Published Online www.clef-campaign.org
12. Perini, M.A.: Para uma Nova Gramática do Português. São Paulo: Ed. Ática (2000)
13. Robertson, S.E., Walker, S.: Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: Proceedings of 17th Annual International ACM SIGIR Conference on Research and Development in IR, pp. 232–241 (1994)
14. www.inf.pucrs.br/gonzalez/TR+
15. www.inf.pucrs.br/gonzalez/TR+/forma
16. www.inf.pucrs.br/gonzalez/TR+/chama
17. Ziviani, N.: Text Operations. In: Baeza-Yates, R., Ribeiro-Neto, B. (eds.) Modern Information Retrieval, ACM Press, New York (1999)

NLP-Driven Constructive Learning for Filtering an IR Document Stream

João Marcelo Azevedo Arcoverde and Maria das Graças Volpe Nunes

Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
Caixa Postal 668, 13560-970 - São Carlos, SP - Brasil
{jmaa,gracan}@icmc.usp.br

Abstract. Feature engineering is known as one of the most important challenges for knowledge acquisition, since any inductive learning system depends upon an efficient representation model to find good solutions to a given problem. We present an NLP-driven constructive learning method for building features based upon noun phrases structures, which are supposed to carry the highest discriminatory information. The method was tested at the CLEF 2006 *Ad-Hoc*, monolingual (Portuguese) IR track. A classification model was obtained using this representation scheme over a small subset of the relevance judgments to filter false-positives documents returned by the IR-system. The goal was to increase the overall precision. The experiment achieved a MAP gain of 41.3%, in average, over three selected topics. The best F1-measure for the text classification task over the proposed text representation model was 77.1%. The results suggest that relevant linguistic features can be exploited by NLP techniques in a domain specific application, and can be used successfully in text categorization, which can act as an important coadjutant process for other high-level IR tasks.

1 Introduction

Modern information access methods have been tightly based on some classification scheme, as the endless growth of indexed material increases the complexity of information retrieval systems. Users expect to retrieve all and only relevant documents according to their needs. Text classification as a complementary filtering task in IR-systems has become a critical issue in knowledge management.

In this paper we study the impact of a linguistically motivated document representation in the context of automatic Text Categorization (TC), as a preliminary filter of other high-level IR activities. An investigation into the feasibility of an NLP-driven constructive learning method was made on the results of CLEF 2006 experiments in the *ad-hoc* retrieval task.

After the relevance judgments for all the search topics were made public, we were able to induce a predictive model over a subset of documents. It is expected that the model can generalize a binary decision for each incoming document

retrieved by the previously ranked results of the IR-task. Only documents that match the similarity estimation must be shown to the user, filtering those that are not relevant, improving average precision over interpolated recall.

In CLEF 2006¹ evaluation for the *ad-hoc*, monolingual (Portuguese) track, our team (NILC²) tested a hybrid indexing scheme that uses statistical and linguistic knowledge based on noun phrases [1]. The method explores both how the hybrid text representation can achieve good effectiveness, and how query expansion could be done in order to produce better recall in detriment of precision.

The selective problem of filtering only the relevant documents from an incoming stream is classificatory, in the sense that there exists a classification criterion to distinguish between two mutually exclusive sets of documents.

The classification problem is the activity of associating a Boolean value for each pair $\langle d_j, c_i \rangle \in D \times C$, where D is the domain of documents, and $C = \{c_1, \dots, c_{|C|}\}$ is the set of predefined categories. The value T associated with $\langle d_j, c_i \rangle$ indicates that the document $d_j \in c_i$, while the value F associated with $\langle d_j, c_i \rangle$ indicates that $d_j \notin c_i$. Hence, a classifier is a function $\Phi: D \times C \rightarrow \{T, F\}$, denoted hypothesis or model, which describes how the documents should be classified.

The present work aims to show how an NLP-driven filtering technique can be applied after the query expansion results, achieving higher evaluation metrics for an IR-system.

Though the predictive model was inferred using a subset of the relevant judgments for each topic of interest, it is accurate to state that if the users knew how to efficiently articulate their information needs through the management of a profile made of positive and negative documents, they would obtain better IR results.

2 Feature Space Representation

Text representation has an important role in information access methods. For information retrieval (IR) and machine learning (ML) systems, there have been many practical and theoretical approaches well described in the literature for text representation, each of which tries to achieve better accurate measures with more or less computational effort. These models vary from unigrams to those that use some dependency strategy between terms (multiterm), or both. The indexing scheme can be obtained by statistical and/or NLP methods.

It is supposed that a suitable multiterm model could augment the precision of concept extraction, even though some experiments show that sophisticated representations may not perform better than unigram models [2]. However, the final word about the use of multiterm models has not been said and there are many studies giving results in favor of multiterm models [3].

We have taken into account that concepts should be well represented by noun phrase structures. These structures constitute a special case of multiterm

¹ www.clef-campaign.org

² nilc.icmc.usp.br

relationship because they carry information with high discriminatory power and informative potential.

The term dependency model adopted in our experiment uses the concept of evidence to weight the relative confidence of its features. The more evidence a feature has within the texts, the higher is its representativity.

3 Feature Weighting Scheme

In order to calculate the feature confidence in a given document of the collection, it is necessary to define a proper weight scheme. The weight of a noun phrase s in a document d follows the Equation (1).

$$w_{s,d} = f_{s,d} \times \sum_{i=1}^n w_{t_i,d} \quad (1)$$

where:

- $f_{s,d}$ is the frequency of occurrence of s in d ;
- $w_{t_i,d}$ is the weight of the i -th term t_i of s in d ; and is defined as:

$$w_{t,d} = \alpha + \beta \times \left(\frac{0.5 \times \log f}{\log f^*} + 0.5 \right) \times \left(\frac{\log \frac{N}{n}}{\log N} \right) \quad (2)$$

which was first introduced by the probabilistic WIN system [4], where:

- $w_{t,d}$ is the weight of term t in d ;
- α is a constant which state that, even if the term does not occur in the document, its probability of relevance isn't zero. The usual value is 0.4, but we used 0 once we obtained better practical results;
- β is the same as $(1 - \alpha)$ and weights the contribution of TF.IDF. The usual value is 0.6, but we used 1;
- f is the frequency of the term t in the document d ;
- f^* is the frequency of the most frequent term in the document;
- N is the size of the collection;
- n is the number of documents containing t

Our IR-system used in CLEF 2006 had a probabilistic model and we wanted to keep compatibility in the feature weighting scheme. As, we needed to satisfy $0 \leq w_{s,d} \leq 1$, and the Equation (1) could not be used, because, given two weights w_1 and w_2 , its sum can exceed 1. Thus, we needed a function that could satisfy the following properties:

- (i) $0 \leq f(w_1, w_2) \leq 1$;
- (ii) $f(w_1, w_2) \geq \max(w_1, w_2)$;

As the weight of a disjunctive term “ s OR t ” in a probabilistic model is given by the Equation (3);

$$1 - [(1 - w_{s,d})(1 - w_{t,d})] \quad (3)$$

and this Equation satisfies the above properties, we have adopted it as a replacement for the sum of weights. Hence, Equation (1) can be converted into the Equation (4).

$$w_{s,d} = 1 - (1 - S)^{f_{s,d}}, \text{ where } S = 1 - \prod_{i=1}^n (1 - w_{t_i,d}) \quad (4)$$

This leads to the final Equation (5).

$$w_{s,d} = 1 - \left(\prod_{i=1}^n (1 - w_{t_i,d}) \right)^{f_{s,d}} \quad (5)$$

4 Experimental Setup

We developed a probabilistic IR-system designed to explore the power of noun phrases as the main descriptors of a hybrid indexing scheme, which aims to combine statistical and linguistic knowledge extracted from the collection. The main objective of this experiment is to conceive a filtering subsystem that acts as a complementary task of other high level IR activities, blocking false-positive documents from the incoming stream returned by the IR-system.

4.1 Collection, Topics and Relevance Judgments

IR-systems are evaluated by constructing a test collection of documents, a set of query topics, and a set of relevance judgments for a subset of the collection (known as pool) with respect to these topics.

For the CLEF 2006 experiment for the *ad-hoc*, monolingual (Portuguese) track, two collections were available, containing documents from Brazil and Portugal. The corpus is composed of four collections: Folha94, Folha95, Publico94 e Publico95³, totalizing 210.736 free text documents (approximately 560Mb).

50 query topics from different subjects were provided, along with their respective descriptions. Each topic represents a long-term user interest, from which it is intended to capture its subjective relevance and to create a search strategy that is supposed to retrieve all and only relevant documents.

The most widely used metrics in IR for the past three decades are precision and recall and the precision-recall curves. However, before they can be computed, it is necessary to obtain the *relevance judgments*, which are a set of positive and negative examples of documents that are supposed to best represent the topic.

³ Complete editions from 1994 and 1995 of journals PÚBLICO (www.publico.pt) and Folha de São Paulo (www.folha.com.br), compiled by Linguateca (www.linguateca.pt).

The *pooling* method was used to generate the CLEF 2006 relevant judgments for this track, which we used as a user's filtering *profile* that acted as a training data set for an automatic Text Classification (TC) task.

4.2 Pre-processing the Collection

The collection required three different levels of pre-processing before indexing. First, the text was segmented in one sentence per line. Then, each term from each sentence was *tokenized* and morphologically analyzed. In this phase we proceed with the morphologic disjunctions, for example, “do = de + o”, “àquele = a + aquele”, “dentre = de + entre”, etc. In the second level of pre-processing, the text was *POS-tagged* using a language modeling proposed by *MXPOST* [5]. In the third level, a TBL-based algorithm [6] was used to identify and tag the noun phrases for each labeled sentence, for each document of the entire collection. It also flagged the nucleus for each noun phrase (which can be formed by multiple lexical words).

It is worth mentioning that the target noun phrases of this experiment are the non-recursive ones, i.e., those that do not hold relative or subordinative clauses. For example, the noun phrase “[o rapaz que subiu no ônibus] é [meu amigo]” (*[the guy that took the bus] is [my friend]*) would be expected to be output as “[o rapaz] que subiu em [o ônibus] é [meu amigo]” (*[the guy] that took [the bus] is [my friend]*). The smallest components of a noun phrase should reflect the most atomic concepts evidenced within the texts, so they cannot give rise to doubts or ambiguities.

5 Constructive Learning

Feature design can demand much human effort depending on the problem addressed. Inadequate description languages or a poorly statistically correlated feature space will certainly intensify the problem, leading to imprecise or excessively complex descriptions. However, these irrelevant or unfit features could be conveniently combined in some context, generating new useful and comprehensible features that could better represent the desired concept. The process of building new features is called Feature Engineering or Constructive Induction (CI) or Constructive Learning [7].

The CI method adopted in this work is automatically and completely conducted by the system. The range of candidate attributes for generating new features is chosen on the analysis of the training data, hence, it is called *data-driven constructive induction* [8]. Those primary constructors define the *problem representation space*, from which we focused the appropriate NLP analysis. This analysis constitutes the basis of the main strategy adopted.

The multiterm feature relevancy is expected to be related to its concept representation, as the feature was constructed from some noun phrase structure, which can be seen as an atomic concept of the real word.

Many attempts have been made to test different strategies to obtain the multiterm features that could best represent the concepts in the documents. In order

to reduce the combinatorial possibilities, we have limited the operators⁴ only to terms which are nuclei of noun phrases, together with their modifiers. By applying some simple heuristics on the operators we build new features, as shown in the following example:

- “o fascinante [cão Chow] de o admirável [doutor Freud]”
- (*the fascinating [dog Chow] of the [admirable Dr. Freud]*)
- uniterm features: fascinante cao chow admiravel doutor freud
- 1st level of multiterm features: cao_chow, doutor_freud
- 2nd level of multiterm features: cao_chow__doutor_freud
- 3rd level of multiterm features: fascinante__cao_chow, admiravel__doutor_freud

There exist three levels for building multiterm features: *i*) a multinuclear NP generates a multiterm feature; *ii*) if a NP has more than one nucleus, the features derived from them are the multiterm features; *iii*) the modifiers of a nucleus are bound to it, producing new features. Sometimes it was necessary to apply some simple heuristics to desambiguate those dependency relationships.

6 Evaluation

A subset of the relevance judgments for each topic was used as training data to induce 50 binary categorizers, one for each CLEF 2006 topic. They were composed, on average, by a small number of non-overlapping and unbalanced positive and negative examples.

The relevant judgments provided an average of 403.08 documents (stdev of 131.49) per topic, divided into 53.54 positives (stdev of 52.25) and 349.54 negatives (stdev of 136.73), in average. The ratio between positive and negative examples is 21.21% (stdev of 24.53%).

It is worth noting that in certain domains the class unbalance problem may cause suboptimal classification performance, as observed in [9]. In such case, the classifier overpowered by the large class tends to ignore the small one. For some topics of our experiment, the negative examples outnumber the positive ones, therefore the bias favors the prediction of false-positive results.

There are “one class learning” approaches, such as the extreme rebalancing [10], which have presented reasonable performance when the unbalanced problem is intrinsic to the domain, and will be investigated for the filtering task in the near future.

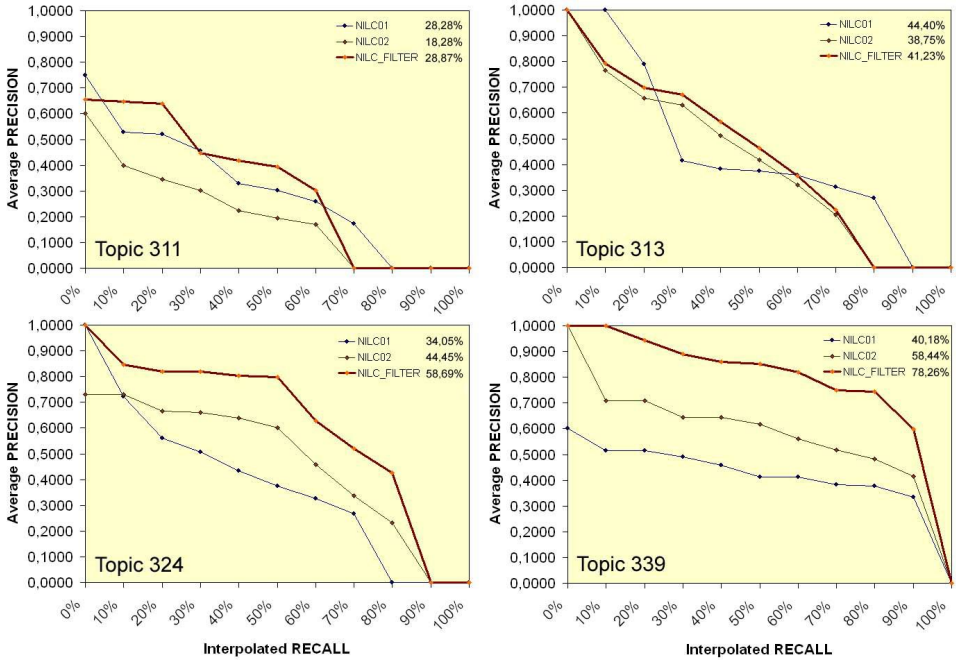
For the experiment presented here, we focused only on those topics for which the classes are balanced. Thus, the evaluation of error-rate and F-measure was carried out over 4 of 50 search topics. They are referred by their numbers: 311, 313, 324 and 339, respectively.

We have used two different families of classifiers to evaluate performance over both uniterm and multiterm document representations: *Naive Bayes* (NB) and *Support Vector Machines* (SVMs).

⁴ Terms that rule as candidates to build new features.

Table 1. F1-measure comparison over classifiers and topics

	SVM_Uniterm	SVM_Multiterm	NB_Uniterm	NB_Multiterm
311	64.35%	69.16%	62.59%	63.20%
313	75.23%	77.10%	70.04%	72.32%
324	71.52%	71.89%	68.36%	69.03%
339	72.59%	75.56%	66.90%	68.37%

**Fig. 1.** Topic MAPs after applying the filter over the NILC02 run

We have used the 10-fold *stratified* cross-validation⁵ to obtain the confusion matrix, and therefore all the related precision, recall, error-rate and F1-measure.

As shown in Table 1, for all the four topics SVMs have shown to be the best classifier over Naive Bayes. Also, the multiterm feature representation provided a slightly better performance over the uniterm, for both classifiers.

The IR-system is evaluated through the MAP (*Mean Average Precision*) metric, which is the mean of the *Average Precision* (AP) over a set of queries, and AP is the average precision after each relevant document is retrieved. In Figure 1 four MAP plots are shown, one for each topic. The runs NILC01 and NILC02 were taken from our CLEF 2006 experiment, representing the initial and expanded

⁵ Preserving same proportion of class distribution among the partitions.

queries. The run NILC_FILTER reveals the filtering result applied over the NILC02 run. For all the four topics the filtering curve achieved the best MAP score.

All the IR and post-filtering experiments were carried out for the multiterm feature representation. We have trained our best classifier (SVM) using only 50% of the relevance judgements, for each topic, preserving the class distribution. We then applied the filter over the same document stream returned by each query topic, which is the same NILC02 run for the expanded query topic used in CLEF 2006. To avoid overfitting and misleading results, the training documents were removed from the stream (test set), so we had to fix the relevance judgment for each topic, reflecting the new scenario without the training data.

7 Conclusion

In this paper an evaluation of the impact of a specific linguistic motivated document representation over TC is reported. We have tested the multiterm feature engineering through a NLP-driven *constructive learning* point-of-view. The TC scenario was used as a post-processing filtering subtask applied over the results of a probabilistic IR-system. From our participation in CLEF 2006 Ad-Hoc, monolingual (Portuguese) language track, we have taken two runs as a baseline to measure the effectiveness of a post-processing filtering activity.

The effective use of NLP in a TC scenario is not associated with a higher accuracy gain over traditional approaches, and requires a substantial increase in computational complexity. Rather, it is related to the improved quality of the produced features, providing a better understanding of the conceptual nature of the categories. This makes it possible to manually manage each category profile vector to aggregate external knowledge. This is certainly useful to extend these systems to a higher level of predictability of any domain-specific content stream.

The TC task was shown to be an important coadjuvant process over other high-level IR activities, improving the overall user experience. In our scenario, the TC acted as a filtering sub-system, blocking spurious content from being delivered to the user with satisfactory precision, once the inductive model had been built under the appropriate circumstances.

References

1. Arcoverde, J.M.A., Nunes, M.d.G.V., Scardua, W.: Using noun phrases for local analysis in automatic query expansion. In: CLEF working notes for ad-hoc, monolingue, Portuguese track (2006)
2. Apté, C., Damerau, F., Weiss, S.M.: Towards language independent automated learning of text categorisation models. In: Research and Development in Information Retrieval, pp. 23–30 (1994)
3. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
4. Turtle, H.R.: Inference Networks for Document Retrieval. PhD thesis (1991)

5. Ratnaparkhi, A.: A maximum entropy part-of-speech tagger, University of Pennsylvania, USA (1996)
6. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4), 543–565 (1995)
7. Michalski, R.S.: Pattern recognition as knowledge-guided computer induction, Tech. Report 927 (1978)
8. Bloedorn, E., Michalski, R.S.: Data-driven constructive induction. *IEEE Intelligent Systems* 13(2), 30–37 (1998)
9. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations* 6(1), 1–6 (2004)
10. Raskutti, B., Kowalczyk, A.: Extreme rebalancing for svms: a case study. *SIGKDD Explorations* 6(1), 60–69 (2004)

ENSM-SE at CLEF 2006 : Fuzzy Proximity Method with an Adhoc Influence Function

Annabelle Mercier and Michel Beigbeder

École Nationale Supérieure des Mines de Saint-Étienne,
158 cours Fauriel, 42023 Saint-Etienne Cedex 2, France
{annabelle.mercier,mbeig}@emse.fr

Abstract. We experiment a new influence function in our information retrieval method that uses the degree of fuzzy proximity of key terms in a document to compute the relevance of the document to the query. The model is based on the idea that the closer the query terms in a document are to each other the more relevant the document. Our model handles Boolean queries but, contrary to the traditional extensions of the basic Boolean information retrieval model, does not use a proximity operator explicitly. A single parameter makes it possible to control the proximity degree required. To improve our system we use a stemming algorithm before indexing, we take a specific influence function and we merge fuzzy proximity result lists built with different width of influence function. We explain how we construct the queries and report the results of our experiments in the ad-hoc monolingual French task of the CLEF 2006 evaluation campaign.

1 Introduction

In the information retrieval domain, systems are based on three basic models: the Boolean model, the vector model and the probabilistic model. These models have many variations (extended Boolean models, models based on fuzzy sets theory, generalized vector space model, . . .) [1]. However, they are all based on weak representations of documents: either sets of terms or bags of terms. In the first case, what the information retrieval system knows about a document is whether it contains a given term or not. In the second case, the system knows the number of occurrences – the *term frequency*, tf – of a given term in each document. So whatever the order of the terms in the documents, they share the same index representation if they use the same terms. Noteworthy exceptions to this rule are most of the Boolean model implementations which propose a NEAR operator [2]. This operator is a kind of AND but with the constraint that the different terms are within a window of size n , where n is an integral value. The set of retrieved documents can be restricted with this operator. For instance, it is possible to discriminate between documents about “data structures” and those about “data about concrete structures”. Using this operator results in an increase in precision of the system [3]. But the Boolean systems that implement a NEAR operator share the same limitation as any basic Boolean system: these systems are not

able to rank the retrieved documents because with this model a document *is* or *is not* relevant to a query. Different extensions have been proposed to the basic Boolean systems to circumvent this limitation. These extensions represent the documents with some kind of term weights. Most of the time these weights are computed on a *tf* basis. Some combining formulas are then applied to compute the document score given the term weights and the query tree. But these extensions are not compatible with the NEAR operator. Some researchers have thus proposed models that attempt to directly score the documents by taking into account the proximity of the query terms within them.

2 Uses of Proximity

Three methods have been proposed to score documents taking into account different sets of intervals containing the query terms. These methods differ in the set of intervals that are selected in a first step, and then in the formulas used to compute the score for a given interval. The method of Clarke et al. [4] selects the shortest intervals that contain all the query terms (this constraint is relaxed if there are not enough retrieved documents), so the intervals cannot be nested. In the method of Hawking et al. [5], for each query term occurrence, the shortest interval containing all the query terms is selected, thus the selected intervals can nest. Rasolofo et al. [6] chose to select intervals only containing *two* terms of the query, but with the additional constraint that the interval is shorter than five words. Moreover, passage retrieval methods indirectly use the notion of proximity. In fact, in several methods, documents are ranked by selecting documents which have passages with a high density of query terms, that is to say documents where the query terms are near to each other [7,8,9]. The next section presents our method which scores documents on the basis of term proximity.

3 Fuzzy Proximity Matching

To address the problem of scoring the documents taking into account the relative order of the words in the document, we have defined a new method based on a *fuzzy proximity* [10] between each position in the document text and a query. This fuzzy proximity function is summed up over to score the document.

We model the fuzzy proximity to an occurrence of a term with an influence function f that reaches its maximum (value 1) at the value 0 and decreases on each side down to 0. Different types of functions (Hamming, rectangular, gaussian, etc.) can be used. We used an adhoc and a triangular one shown in Figure 1. In the following, the examples and the experiments will be based on a triangular function $x \mapsto \max(\frac{k-|x|}{k}, 0)$. The constant k controls the support of the function and this support represents the extent of influence of each term occurrence. A similar parameter can be found for other shapes.

So, for a query term t , the fuzzy proximity function to the occurrence at position i of the term t is $x \mapsto f(x - i)$. Now, we define the term proximity

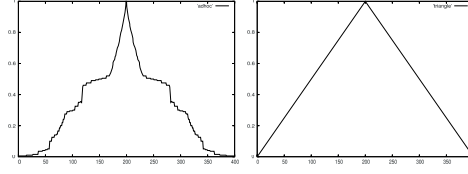


Fig. 1. The adhoc and triangular influence functions used in experiments

function w_t^d which models the fuzzy proximity at the position x in the text to the term t by combining the fuzzy proximity functions of the different occurrences of the term t :

$$x \mapsto w_t^d(x) = \max_{i \in Occ(t,d)} f(x - i)$$

where $Occ(t, d)$ is the set of the positions of the term t in the document d and f is the influence function.

The query model is the classical Boolean model: A tree with terms on the leaves and OR or AND operators on the internal nodes. At an internal node, the proximity functions of the sons of this node are combined in the query tree with the usual fuzzy set theory formulas. So the fuzzy proximity is computed by

$$w_{q \text{ OR } q'}^d = \max(w_q^d, w_{q'}^d)$$

for a disjunctive node and by

$$w_{q \text{ AND } q'}^d = \min(w_q^d, w_{q'}^d)$$

for a conjunctive node. With a post-order tree traversal a fuzzy proximity function to the query can be computed at the root of the query tree as the fuzzy proximity functions are defined on the leaves.

So we obtain a function w_q^d from x to the interval $[0, 1]$. The result of the summation of this function is used as the score of the document:

$$s(q, d) = \sum_{x=-\infty}^{+\infty} w_q^d(x) \ .$$

Thus, the computed score $s(q, d)$ depends on the fuzzy proximity functions and enables document ranking according to the query term proximity in the documents.

4 Experiments and Evaluation

We carried out experiments within the context of the CLEF 2006 evaluation campaign in the ad-hoc monolingual French task¹. We used the retrieval search engine LUCY² which is based on the Okapi information retrieval model [11] to

¹ <http://clef.isti.cnr.it/>

² <http://www.seg.rmit.edu.au/lucy/>

index this collection. It was easy to adapt this tool to our method because it keeps the positions of the terms occurring in the documents in the index. Thus, we extended this tool to compute the relevance score values for our fuzzy proximity matching function.

Documents in the CLEF 2006 test collection are newspapers articles in XML format from *SDA* and *Le Monde* of the years 1994 and 1995. For each document (tag <DOC>), we keep the fields <DOCNO> with the tag and the document number, the textual contents of the tags <TX>, <LD>, <TI>, <ST> for *SDA French* and <TEXT>, <LEAD1>, <TITLE> for *Le Monde* 1995. We used the topics and the relevance judgements to evaluate the different methods by the `trec_eval` program.

4.1 Building the Queries

Each topic is composed of three tags: <FR-title>, <FR-desc>, <FR-narr>. Two sets of queries were built for our experiments.

Automatically built queries. For this set, a query is built with the terms from the title field where the stop words³ are removed. Here is an example with the topic #278. The original topic is expressed by:

```
<top>
<num> 278 </num>
<FR-title> Les moyens de transport pour handicapés</FR-title>
<FR-desc> A quels problèmes doivent faire face les personnes
handicapées physiques lorsquelles empruntent les transports
publics et quelles solutions sont proposées ou adoptées?
</FR-desc>
<FR-narr> Les documents pertinents devront décrire les
difficultés auxquelles doivent faire face les personnes
diminuées physiquement lorsquelles utilisent les transports
publics et/ou traiter des progrès accomplis pour résoudre ces
problèmes.
</FR-narr>
</top>
```

First, the topic number and the title field are extracted and concatenated:

```
278 moyens transport handicapés
```

From this form, the queries are automatically built by simple derivations:

```
Lucy:                278 moyens transport handicapés
conjunctive fuzzy proximity: 278 moyens & transport & handicapés
disjunctive fuzzy proximity: 278 moyens | transport | handicapés
```

³ Removed stop words: à, aux, au, chez, et, dans, des, de, du, en, la, les, le, par, sur, uns, unes, une, un, d', l'.

Queries for the method implemented in the LUCY tool are flat queries composed of different inflectional and/or derivational forms of the terms. Here is an example for topic #278:

fuzzy proximity: 278 (moyen | moyens) & (transport | transports)
 & (handicap | handicapé | handicapés)
 Lucy: 278 moyen moyens transport transports
 handicap handicapé handicapés

The Okapi model and our fuzzy method were compared. It is well known that the Okapi method gives one of the best performances. However, a previous study showed that proximity based methods improve retrieval [12]. If one of our experiments with our proximity based method does not retrieve enough documents (one thousand for the CLEF experiments), then its results list is supplemented by documents from the Okapi result list that have not yet been retrieved by the proximity based method.

We note in past experiments that the higher the area ($k = 200$ was used) of influence of a term the better the results are. Moreover, we retrieved more documents with fuzzy proximity with a large width of influence function. So, we merge for each type of queries (title, description and manual), the results obtained with several k values (k equal to 200, 100, 80, 50, 20, 5).

In the official runs, the queries used with fuzzy proximity method and adhoc influence function were:

1. the conjunction of the terms automatically extracted from the title field in run **RIMAM06TL**;
2. the conjunction of the terms automatically extracted from the description field in run **RIMAM06TDNL** and;
3. manually built queries with terms from the three fields run **RIMAM06TDML**.

Here we present with CLEF 2005 queries the differences obtained between runs with stemming (or not) at indexing step or at query formulation step.

For the runs where the Okapi method was used, the queries are flat (bag of terms). These runs were produced by using the native LUCY search engine

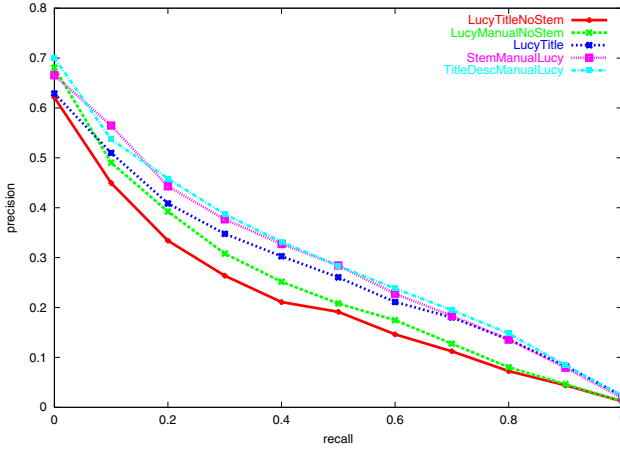


Fig. 2. Okapi Lucy - Result with stemming at indexing (LucyTitle, StemManualLucy, TitleDescManualLucy) and no stemming (LucyTitleNoStem, LucyManualNoStem)

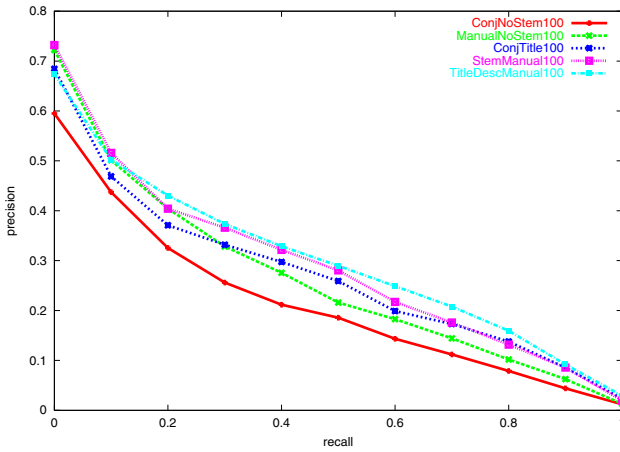


Fig. 3. Fuzzy Proximity with $k = 100$ - Result with stemming at indexing (ConjTitle100, StemManual100, TitleDescManual100) and no stemming (ConjNoStem100, ManualNoStem100).

and they provide the baselines for the comparison with our method. The recall precision results are provided in Figure 2.

We can see that the runs with no stemming step before indexing have less precision than the others. Figure 3 also shows that the stemming step provide better results. But we can see that the run TitleDescManual is above the ConjTitle100 which means that the words added for “stemming” queries increase the precision results. Figure 4 shows the differences between the Okapi Lucy method and the fuzzy proximity method. We can see that our method is better

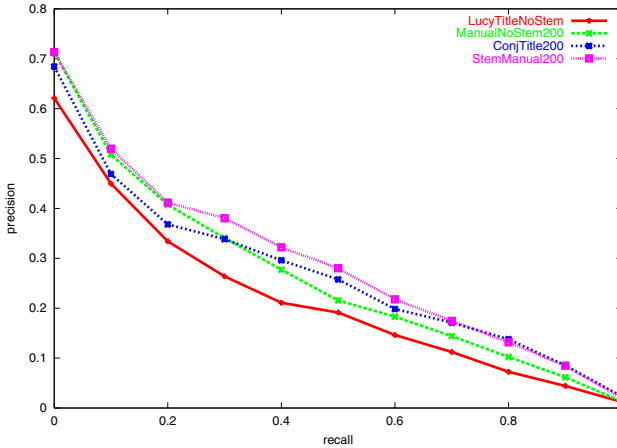


Fig. 4. Fuzzy Proximity with $k = 100$ - Result with stemming at indexing (ConjTitle200, StemManual200) and no stemming (LucyTitleNoStem, ManualNoStem200)

than the Lucy one (ManualNoStem200 vs. LucyTitleNoStem; StemManual200 vs. ConjTitle200). We can note the run with stemming at indexing and at query time is the best.

5 Conclusion

We have presented our information retrieval model which takes into account the position of the query terms in the documents to compute the relevance scores. We experimented this method on the CLEF 2006 Ad-Hoc French test collection. In these experiments, we submit runs which use different k values in order to retrieve more documents with our method. We plan to study in detail the fusion of result lists and explore the idea to merge results from runs built with different values of parameter k . Another idea is to adapt the value of this parameter with respect to the document size and then introduce a kind of normalisation.

The main results of our CLEF experiments regard the comparison between stemming treatment at the indexing time or stemming simulation at query time : a detailed analysis of this work is available in [10] where we show the comparative benefits of stemming in the two phases. The results of this study lead us to think that a human-assisted method can be developed to build boolean queries. Perhaps, the user should add query words suggested by a thesaurus in order to complete the boolean queries and retrieve more documents with our method. In our next experiments, we plan to adapt the fuzzy proximity method to structured documents and apply this to the CLEF collection by using the title field and the text field of the XML documents differently.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. ACM Press / Addison-Wesley, New York (1999)
2. Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York (1983)
3. Keen, E.M.: Some aspects of proximity searching in text retrieval systems. *Journal of Information Science* 18, 89–98 (1992)
4. Clarke, C.L.A., Cormack, G.V., Tudhope, E.A.: Relevance ranking for one to three term queries. *Information Processing and Management* 36(2), 291–311 (2000)
5. Hawking, D., Thistlewaite, P.: Proximity operators - so near and yet so far. In: Harman, D.K. (ed.) *The Fourth Text REtrieval Conference (TREC-4)*, Department of Commerce, National Institute of Standards and Technology, pp. 131–143 (1995)
6. Rasolofo, Y., Savoy, J.: Term proximity scoring for keyword-based retrieval systems. In: Sebastiani, F. (ed.) *ECIR 2003*. LNCS, vol. 2633, pp. 207–218. Springer, Heidelberg (2003)
7. Wilkinson, R.: Effective retrieval of structured documents. In: *SIGIR '94, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 311–317. Springer, New York (1994)
8. De Kretser, O., Moffat, A.: Effective document presentation with a locality-based similarity heuristic. In: *SIGIR '99: Proceedings of the 22nd ACM SIGIR Annual International Conference on Research and Development in Information Retrieval*, pp. 113–120. ACM, New York (1999)
9. Kise, K., Junker, M., Dengel, A., Matsumoto, K.: Passage retrieval based on density distributions of terms and its applications to document retrieval and question answering. In: Dengel, A., Junker, M., Weisbecker, A. (eds.) *Reading and Learning*. LNCS, vol. 2956, pp. 306–327. Springer, Heidelberg (2004)
10. Mercier, A.: *Modélisation et prototypage d'un système de recherche d'informations basé sur la proximité des occurrences de termes de la requête dans les documents*. Ph.d dissertation, Ecole Nationale Supérieure des Mines de Saint Etienne, Centre G2I (2006)
11. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at trec-3. In: Harman, D.K. (ed.) *Overview of the Third Text REtrieval Conference (TREC-3)*, Department of Commerce, National Institute of Standards and Technology, pp. 109–126 (1994)
12. Mercier, A.: Étude comparative de trois approches utilisant la proximité entre les termes de la requête pour le calcul des scores des documents. In: *INFORSID 2004*, pp. 95–106 (2004)

A Study on the Use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval

Viviane Moreira Orengo, Luciana S. Buriol, and Alexandre Ramos Coelho

UFRGS, Instituto de Informática, Porto Alegre, Brazil
{vmorengo, buriol, arcoelho}@inf.ufrgs.br

Abstract. For UFRGS's first participation in CLEF our goal was to compare the performance of heavier and lighter stemming strategies using the Portuguese data collections for monolingual Ad-hoc retrieval. The results show that the safest strategy was to use the lighter alternative (reducing plural forms only). On a query-by-query analysis, full stemming achieved the highest improvement but also the biggest decrease in performance when compared to no stemming. In addition, statistical tests showed that the only significant improvement in terms of mean average precision, precision at ten and number of relevant retrieved was achieved by our lighter stemmer.

1 Introduction

This paper reports on monolingual information retrieval experiments that we have performed for CLEF2006. We took part in the ad-hoc monolingual track, focusing on the Portuguese test collections.

Our aim was to compare the performance of lighter and heavier stemming alternatives. We compared two different algorithms: a Portuguese version of the Porter stemmer and the “Removedor de Sufixos da Língua Portuguesa” (RSLP) [4]. Moreover, a simpler version of the RSLP stemmer that reduces plural forms only was also tested. We compared the results of the three stemming alternatives with results of no stemming.

The remainder of this paper is organised as follows: Section 2 presents the RSLP stemmer; Section 3 discusses the experiments and results; and Section 4 presents the conclusions.

2 The Stemming Algorithm

We have used the RSLP algorithm, proposed in our earlier work [4]. It was implemented in C and is freely available from <http://www.inf.ufrgs.br/~vmorengo/rslp>.

This section introduces the algorithm. RSLP is based solely on a set of rules (not using any dictionaries) and is composed by 8 steps that need to be executed in a certain order. Figure 1 shows the sequence those steps must follow:

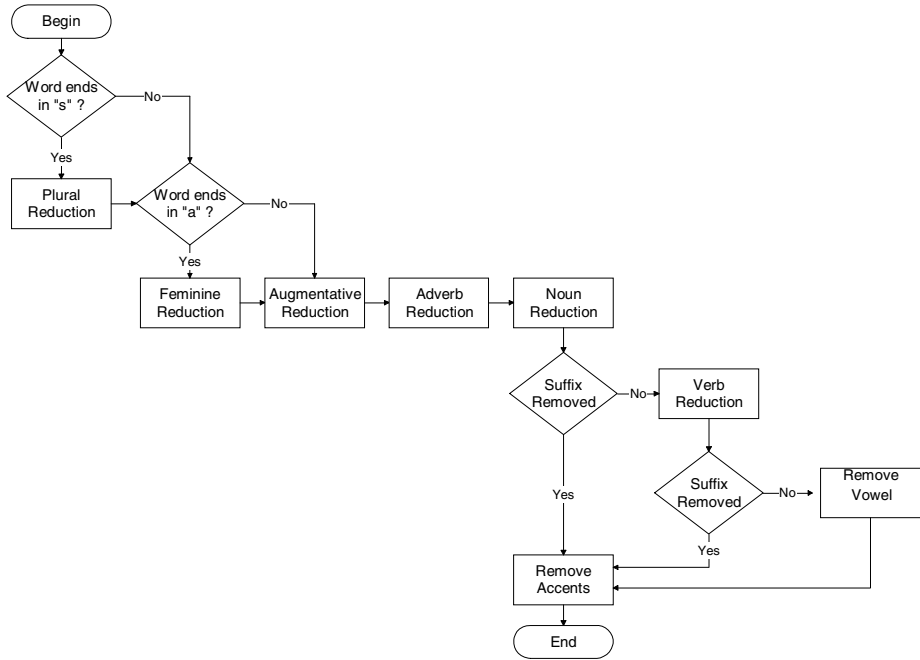


Fig. 1. Sequence of Steps for the RSLP Algorithm

Each step has a set of rules, the rules in the steps are examined in sequence and only one rule in a step can apply. However, the same word might be stemmed by more than one step. For example *francesinhas* (little French girls), would first go through the plural reduction step and become *francesinha*. Next, it would be stemmed by the feminine step, becoming *francesinho*. Finally, the augmentative/diminutive step would produce *frances*.

The longest possible suffix is always removed first because of the ordering of the rules within a step, e.g. the plural suffix *-es* should be tested before the suffix *-s*. At the moment, the Portuguese Stemmer contains 253 rules. please refer to web page for the complete list.

Each rule states:

- The *suffix* to be removed;
- The *minimum length* of the stem: this is to avoid removing a suffix when the stem is too short. This measure varies for each suffix, and the values were set by observing lists of words ending in the given suffix. Although there is no linguistic support for this procedure it reduces overstemming errors. Overstemming is the removal of a sequence of characters that is part of the stem and not a suffix.
- A *replacement suffix* to be appended to the stem, if applicable;

- A *list of exceptions*: for nearly all rules we defined, there were exceptions, so we added exception lists for each rule. Such lists were constructed with the aid of a vocabulary of 32,000 Portuguese words freely available from [6]. Tests with the stemmer have shown that exceptions list reduce overstemming errors by 5%.

An example of a rule is:

```
"inho", 3, "", {"caminho", "carinho", "cominho", "golfinho",  
"padrinho", "sobrinho", "vizinho"}
```

Where *inho* is a suffix that denotes diminutive, 3 is the minimum size for the stem, which prevents words like *linho* (linen) from being stemmed and the words between brackets are the exceptions for this rule, that is, they end in the suffix but they are not diminutives. All other words that end in *-inho* and that are longer than 6 characters will be stemmed. There is no replacement suffix in this rule. Below we explain the eight steps involved in our stemming procedure.

Step 1: Plural Reduction

With rare exceptions, the plural forms in Portuguese end in *-s*. However, not all words ending in *-s* denote plural, e.g. *lápiz* (pencil). This step consists basically in removing the final *s* of the words that are not listed as exceptions. Yet sometimes a few extra modifications are needed e.g. words ending in *-ns* should have that suffix replaced by *m* like in *bons* → *bom*. There are 11 stemming rules in this step.

Step 2: Feminine Reduction

All nouns and adjectives in Portuguese have a gender. This step consists in transforming feminine forms to their corresponding masculine. Only words ending in *-a* are tested in this step but not all of them are converted, just the ones ending in the most common suffixes, e.g. *chinesa* → *chinês*. This step is composed by 15 rules.

Step 3: Adverb Reduction

This is the shortest step of all, as there is just one suffix that denotes adverbs *-mente*. Again not all words with that ending are adverbs so an exception list is needed.

Step 4: Augmentative/Diminutive Reduction

Portuguese nouns and adjectives present far more variant forms than their English counterparts. Words have augmentative, diminutive and superlative forms e.g. "small house" = *casinha*, where *-inha* is the suffix that indicates a diminutive. Those cases are treated by this step. According to Cunha & Lindley-Cintra [1], there are 38 of these suffixes. However, some of them are obsolete therefore, in order to avoid overstemming, our algorithm uses only the most common ones that are still in usage. This step comprises 23 rules.

Step 5: Noun Suffix Reduction

This step tests words against 84 noun (and adjective) endings. If a suffix is removed here, steps 6 and 7 are not executed.

Step 6: Verb Suffix Reduction

Portuguese is a very rich language in terms of verbal forms, while the regular verbs in English have just 4 variations (e.g. talk, talks, talked, talking), the Portuguese regular verbs have over 50 different forms [3]. Each one has its specific suffix. The verbs can vary according to tense, person, number and mode. The structure of the verbal forms can be represented as: root + thematic vowel + tense + person, e.g. *and + a + ra + m* (they walked). Verbal forms are reduced to their root, by 101 stemming rules.

Step 7: Vowel Removal

This task consists in removing the last vowel (“a”, “e” or “o”) of the words which have not been stemmed by steps 5 and 6, e.g. the word *menino* (boy) would not suffer any modifications by the previous steps, therefore this step will remove its final -o, so that it can be conflated with other variant forms such as *menina*, *meninice*, *meninão*, *menininho*, which will also be converted to the stem *menin*.

Step 8: Accents Removal

Removing accents is necessary because there are cases in which some variant forms of the word are accented and some are not, like in *psicólogo* (psychologist) and *psicologia* (psychology), after this step both forms would be conflated to *psicolog*. It is important that this step is done at this point and not right at the beginning of the algorithm because the presence of accents is significant for some rules e.g. *is* → *ol* transforming *sóis* (suns) to *sol* (sun). If the rule was *ois* → *ol* instead, it would make mistakes like stemming *dois* (two) to *dol*. There are 11 rules in this step.

It is possible that a single input word needs to go through all stemming steps. In terms of time complexity this represents the worst case possible.

The Portuguese version of the Porter Stemmer and the RSLP are based solely on rules that need to be applied in a certain order. However, there are some differences between the two stemmers:

- The number of rules - RSLP has many more rules than the Portuguese Porter because it was designed specifically for Portuguese. There are some morphological changes such as augmentatives and feminine forms that are not treated by the Portuguese Porter Stemmer.
- The use of exceptions lists - RSLP includes a list of exceptions for each rule as they help reducing overstemming errors.
- The steps composing the two algorithms are different.

3 Experiments

This section describes our experiments submitted to the CLEF 2006 campaign. Section 3.1 details the resources used, Section 3.2 presents the results and Section 3.3 shows the statistics of the stemming process.

3.1 Description of Runs and Resources

The Portuguese data collections were indexed using SMART [5]. We used the title and description fields of the query topics. Query terms were automatically extracted from the topics. Stop words were removed from both documents and topics. In addition, terms such as “find documents” were removed from the topics.

The processing time was less than 4 minutes for all runs in a Sunfire V240 with 6 GB of main memory and 2 ultra sparc IIIi processors of 1Ghz, under SunOS 5.9. This includes indexing the 210,734 documents and running all 50 queries.

Four runs were tested:

- **NoStem** - No stemming was applied, this run was used as the baseline
- **Porter** - Full stemming using the Portuguese version of the Porter stemmer
- **RSLP** - Full stemming using the RSLP stemmer
- **RSLP-S** - applying only the first step of RSLP to deal with plural reduction only

3.2 Results

Table 1 shows the number of terms indexed in each run. Full stemming with RSLP achieved the highest reduction on the number of entries, followed by the Portuguese version of the Porter stemmer. The lighter stemming strategy reduced the number of entries by 15%.

Table 1. Number of Terms in the Dictionary for all runs. The percentages indicate the reduction attained by each stemming procedure in relation to the baseline.

Run	Number of Terms
NoStem	425996
Porter	248121 (-41.75%)
RSLP	225356 (-47.10%)
RSLP-S	358299 (-15.89%)

The results show that the best performance, in terms of mean average precision (MAP), was achieved by RSLP-S. Both runs in which full stemming was performed achieved identical results in terms of MAP. However, the RSLP outperformed the Portuguese version of the Porter stemmer in terms of Pr@10, but the difference was only marginal. In terms of relevant documents retrieved, a *T-test* has shown that both Porter and RSLP-S have a significantly retrieved a larger number of relevant documents.

In order to see whether the performance improvements shown in Table 1 are statistically significant, a paired *T-test* was performed. Although our data is not

Table 2. Results in terms of MAP, Pr@10 and average number of relevant documents retrieved. The asterisk denotes a statistically significant improvement in relation to the baseline.

Run	Mean Average Precision	Precision at 10	Avg of Rel Ret
NoStem	0.2590	0.3880	38.70
Porter	0.2790 (+7.72%)	0,4260 (+9.79%)	43.12*
RSLP	0.2790 (+7.72%)	0,4320 (+11.34%)	42.18
RSLP-S	0.2821 (+8.91%)*	0,4300 (+10.82%)*	42.48*

perfectly normally distributed, Hull [2] argues that the *T-test* performs well even in such cases. The standard threshold for statistical significance (α) of 0.05 was used. When the calculated p value is less than α , there is a significant difference between the two experimental runs. The results of the statistical tests show that full stemming does not produce a statistically significant improvement (in terms of both MAP and Pr@10) for either algorithm (p values of 0.25 for RSLP and 0.22 for Porter considering MAP and p values of 0.14 for RSLP and 0.18 for Porter when analysing Pr@10). Porter has achieved a significant improvement in terms of number of relevant documents retrieved (p value = 0.04). RSLP-S, however, has achieved a statistically significant improvement compared to baseline for MAP, Pr@10 and number of relevant documents retrieved (p values of 0.003, 0.01 and 0.04 respectively). Figure 2 shows recall-precision curves for all runs.

A query-by-query analysis, shown in Table 3, demonstrates that for 12 topics no stemming was the best alternative. Some form of stemming helped 38 out of 50 topics. Confirming the results in terms of MAP and Pr@10, the best per-

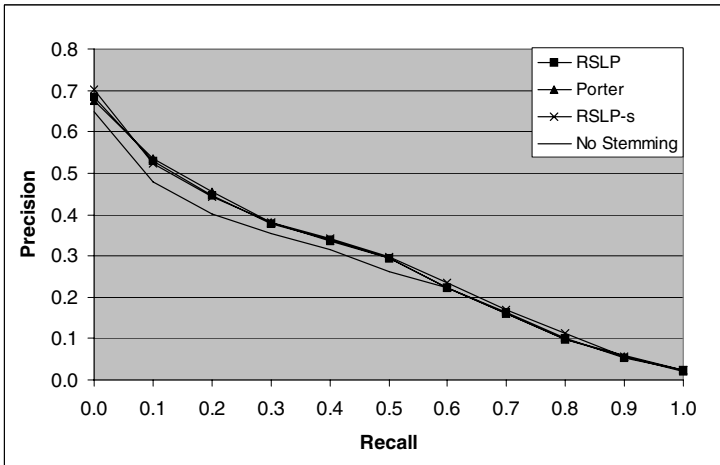


Fig. 2. Recall-precision curves for all runs

formance was achieved by the lighter stemming alternative RSLP-S. Full stemming with RSLP achieved the biggest performance improvement (topic 340 AvP 0.0003 $\rightarrow$ 0.3039), but also the biggest drop (topic 343 AVP 0.4276 $\rightarrow$ 0.1243). Stemming also helped finding 221 relevant documents that were not retrieved by the NoStem run.

Table 3. Runs and the number of topics in which they achieved the best average precision

Run	Number of Topics
NoStem	12
Porter	10
RSLP	12
RSLP-S	16
Total	50

It seemed plausible that queries with few relevant documents would benefit more from stemming, resulting in a negative correlation between the number of relevant documents for the topic and the change in performance achieved with stemming. However, a weak positive correlation of 0.15 was found. We would like to be able to predict the types of queries that would be benefited from stemming, but that needs further analysis with a larger number of topics.

Table 4. Percentage of Rule Applications by step

Step	% of Applications
Plural	17.55%
Feminine	5.08%
Adverb	0.82%
Augmentative	5.17%
Noun	21.01%
Verb	15.78%
Vowel	34.59%

3.3 RSLP Statistics

In this section we report on the statistics generated by the processing of the RSLP stemmer. The stemmer has processed a total of 53,047,190 words. The step with the largest number of applications is Vowel Removal (step 7), followed by Noun (step 5) and Verb (step 6). The least applied step was Adverb (step 3). The rules for Noun, Verb, Plural and Vowel account for approximately 90% of all reductions. Table 4 shows the percentage of rule applications by step. We have noticed that 11 rules (4.5%) of the rules were never applied. Those rules correspond to rare suffixes that were not present in the text.

4 Conclusions

This paper reported on monolingual ad-hoc IR experiments using Portuguese test collections. We evaluated the validity of stemming comparing the Portuguese version of the Porter stemmer and two versions of the RSLP stemmer, one that applies full stemming and one that only reduces plural forms. Below we summarise our conclusions:

- The lighter version of the RSLP stemmer yields statistically significant performance improvements in terms of MAP, Pr@10 and number of relevant retrieved.
- Full stemming, both with Porter and RSLP, has improved the results in terms of MAP, Pr@10 and Relevant Retrieved. However the difference was not statistically significant.
- On a query-by-query analysis we found that stemming helped 38 out of 50 topics and that it enabled the retrieval of 221 further relevant documents that were missed by the run in which no stemming was used.

We can conclude that lighter stemming is the most beneficial strategy for monolingual ad-hoc Portuguese retrieval. This is due to the fact that heavy stemming might introduce noise in the retrieval process and thus harm the performance for some query topics.

Acknowledgements

This work was supported by a CAPES-PRODOC grant from the Brazilian Ministry of Education.

References

1. Cunha, C., Lindley-Cintra, L.: *Nova Gramática do Português Contemporâneo*. Rio de Janeiro: Nova Fronteira (in Portuguese, 1985)
2. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval experiments. In: *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 329–338. Pittsburgh (1993)
3. Macambira, J.R.: *A Estrutura Morfo-Sintática do Português*. São Paulo, Brazil: Ed. Pioneira (in Portuguese, 1999)
4. Orengo, V.M., Huyck, C.R.: A Stemming Algorithm for the Portuguese Language. In: *8th International Symposium on String Processing and Information Retrieval (SPIRE)*. Laguna de San Raphael, Chile, pp. 183–193 (2001)
5. SMART Information Retrieval System, <ftp://ftp.cs.cornell.edu/pub/smart/>
6. Snowball, <http://snowball.tartarus.org/portuguese/voc.txt>

Benefits of Resource-Based Stemming in Hungarian Information Retrieval

Péter Halácsy¹ and Viktor Trón²

¹ Budapest University of Technology and Economics
Centre for Media Research
`hp@mokk.bme.hu`

² International Graduate College
Saarland University and University of Edinburgh
`v.tron@ed.ac.uk`

Abstract. This paper discusses the impact of resource-driven stemming in information retrieval tasks. We conducted experiments in order to identify the relative benefit of various stemming strategies in a language with highly complex morphology. The results reveal the importance of various aspects of stemming in enhancing system performance in the IR task of the CLEF ad-hoc monolingual Hungarian track.

The first Hungarian test collection for information retrieval (IR) appeared in the 2005 CLEF ad-hoc task monolingual track. Prior to that no experiments had been published that measured the effect of Hungarian language-specific knowledge on retrieval performance.

Hungarian is a language with highly complex morphology.¹ Its inventory of morphological processes include both affixation (prefix and suffix) and compounding. Morphological processes are standardly viewed as providing the grammatical means for (i) creating new lexemes (derivation) as well as (ii) expressing morphosyntactic variants belonging to a lexeme (inflection). To illustrate the complexity of Hungarian morphology, we mention that a nominal stem can be followed by 7 types of Possessive, 3 Plural, 3 Anaphoric Possessive and 17 Case suffixes yielding as many as 1134 possible inflected forms.

Similarly to German and Finnish, compounding is very productive in Hungarian. Almost any two (or more) nominals next to each other can form a compound (e.g., *üveg+ház+hat-ás* = glass+house+effect 'greenhouse effect'). The complexity of Hungarian and the problems it creates for IR is detailed in [1].

1 Stemming and Hungarian

All of the top five systems of the 2005 track (Table 1) had some method for handling the rich morphology of Hungarian: either words were tokenized to n -grams or an algorithmic stemmer was used.

¹ A more detailed descriptive grammar of Hungarian is available at <http://mokk.bme.hu/resources/ir>

Table 1. The top five runs for Hungarian ad hoc monolingual task of CLEF 2005

part	run	map	stemming method
jhu/apl	aplmohud	41.12%	4gram
unine	UniNEhu3	38.89%	Savoy's stemmer + decompounding
miracle	xNP01ST1	35.20%	Savoy's stemmer
humminngbird	humHU05tde	33.09%	Savoy's stemmer + 4gram
hildesheim	UHIHU2	32.64%	5gram

The best result was achieved by JHU/APL with an IR system based on language modelling in the run called *aplmohud* [2]. This system used a character 4-gram based tokenization. Such n -gram techniques can efficiently get round the problem of rich agglutinative morphology and compounding. For example the word *atomenergia* = 'atomic energy' in the query is tokenized to *atom*, *tome*, *omen*, *mene*, *ener*, *nerg*, *ergi*, *rgia* strings. When the text only contains the form *atomenergiával* = 'with atomic energy', the system still finds the relevant document.

Although this system used the Snowball stemmer together with the n -gram tokenization for the English and French tasks, the Hungarian results were nearly as good: English 43.46%, French 41.22% and Hungarian 41.12%. From these results it seems that the difference between the isolating and agglutinating languages can be eliminated by character n -gram methods.

Unine [3], Miracle [4] and Hummingbird [5] all employ the same algorithmic stemmer for Hungarian that removes the nominal suffixes corresponding to the different cases, the possessive and plural (<http://www.unine.ch/info/clef>).

UniNEhu3 [3] also uses a language independent decompounding algorithm that tries to segment words according to corpus statistics calculated from the document collection [6]. The idea is to find a segmentation that maximizes the probability of hypothesized segments given the document and the language. Given the density of short words in the language, spurious segmentations can be avoided by setting a minimum length limit (8-characters in the case of Savoy) on the words to be segmented.

Among the 2005 CLEF contestants, [7] is especially important for us, since she uses the same baseline system. She developed four stemmers which implement successively more aggressive stripping strategies. The lightest only strips some of the frequent case suffixes and the plural and the heaviest strips all major inflections. We conjecture that the order in which the suffix list is enriched is based on an intuitive scale of suffix transparency or meaningfulness which is assumed to impact on the retrieval results. In line with previous findings, she reports that the stemming enhances retrieval, with the most aggressive strategy winning. However, the title of her paper 'Four stemmers and a funeral' aptly captures her main finding that even the best of her stemmers performs the same as a 6-gram method.

2 The Experimental Setting

In order to measure the effects of various flavours of stemming on retrieval performance, we put together an unsophisticated IR system. We used JAKARTA LUCENE 2.0, an off-the-shelf system to perform indexing and retrieval with ranking based on its default vector space model. No further ranking heuristics or post-retrieval query expansion is applied. Before indexing the XML documents are converted to plain text format, header (title, lead, source, etc.) and body sequentially appended. For tokenization we used Lucene's `LetterTokenizer` class: this considers every non-alphanumeric character (according to the Java `Character` class) as token boundary. All tokens are lowercased but not stripped of accents. The document and query texts are tokenized the same way including topic and description fields used in the search. Our various runs differ only in how these text tokens are mapped on terms for document indexing and querying. In the current context, we define stemming as solutions to this mapping. Tokens that exactly matched a stopword² before or after the application of our stemming algorithms were eliminated. Stemming can map initial document and query tokens onto zero, one or more (zero only for stopwords) terms. If stemming yields more than one term, each resulting term (but not the original token) was used as an index term. For the construction of the query each term resulting from the stemmed token (including multiple ones) was used as a disjunctive query term.

Using this simple framework allows us to isolate and compare the impact of various strategies of stemming. Using an unsophisticated architecture has the further advantage that any results reminiscent of a competitive outcome will suggest the beneficial effect of the particular stemming method even if no direct comparison to other systems is available due to different retrieval and ranking solution employed.

2.1 Strategies of Stemming

Instead of some algorithmic stemmers employed in all previous work, we leveraged our word-analysis technology designed for generic NLP tasks including morphological analysis and POS-tagging. **Hunmorph** is a word-analysis toolkit, with a language independent analyser using a language-specific lexicon and morphological grammar [8]. The core engine for recognition and analysis can (i) perform guessing (i.e., hypothesize new stems) and (ii) analyse compounds.

Guessing means that possible analyses (morphosyntactic tag and hypothesized lemma) are given even if the input word's stem is not in the lexicon. This feature allows for a stemming mode very similar to resourceless algorithmic stemmers if no lexicon is used. However, guessing can be used in addition to the lexicon.

To facilitate resource sharing and to enable systematic task-dependent optimizations from a central lexical knowledge base, the toolkit offers a general framework for describing the lexicon and morphology of any language. **hunmorph** uses the Hungarian lexical database and morphological grammar called

² We used the same stopword list as [7], which is downloadable at <http://ilps.science.uva.nl/Resources/HungarianStemmer/>

`morphdb.hu` [9]. `morphdb.hu` is by far the broadest-coverage resource for Hungarian reaching about 93% recall on the 700M word Hungarian Webcorpus [10]. We set out to test the impact of using this lexical resource and grammar in an IR task. In particular we wanted to test to what extent guessing can compensate for the lack of the lexicon and what types of affixes should be recognized (stripped). We also compared this technology with the two other stemmers mentioned above, Savoy’s stemmer and Snowball.

Decompounding is done based on the compound analyses of `hunmorph` according to compounding rules in the resource. In our resource, only two nominals can form a compound. Although compounding can be used with guessing, this only makes sense if the head of the compound can be hypothesized independently, i.e., if we use a lexicon. We wanted to test to what extent compounding boosts IR efficiency.

Due to extensive morphological ambiguities, `hunmorph` often gives several alternative analyses. Due to limitations of our simple IR system we choose only one candidate. We have various strategies as to which of these alternatives should be retained as the index term. We can use (i) basic heuristics to choose from the alternants, or (ii) use a POS-tagger that disambiguates the analysis based on textual context.

As a general heuristics used in (i) and (ii) we prefer analyses that are neither compound nor guessed, if no such analysis exists then we prefer non-guessed compounds over guessed analyses.

(ii) involves a linguistically sophisticated method, pos-tagging that restricts the candidate set by choosing an inflectional category based on contextual disambiguation. We used a statistical POS-tagger [11] and tested its impact on IR performance. This method relies on the availability of a large tagged training corpus; if a language has such a corpus, lexical resources for stemming are very likely to exist. Therefore it seemed somewhat irrelevant to test the effect of pos-tagging without lexical resources (only on guesser output).³

If there are still multiple analyses either with or without POS-tagging, we found that choosing the shortest lemma for guessed analyses (aggressive stemming) and the longest lemma for known analyses (blocking of further analysis by known lexemes) works best. When a compound analysis is chosen, lemmata of all constituents are retained as index terms.

3 Evaluation

Table 2 compares the performance of the various stemming algorithms. The clearest conclusion is the robust increase in precision achieved when stemming is used. Either of the three stemmers, Savoy, Snowball and Hunmorph is able to boost precision with at least 50% in both years. Pairwise comparisons with a paired t-test on topic-wise map values show that all three stemmers are significantly better than the baseline. Snowball and Savoy are not different. Most

³ POS-tagging also needs sentence boundary detection. For this we trained a maximum entropy model on the Hungarian Webcorpus [12].

Table 2. Results of different stemming methods

stemmer	year	MAP	P10	ret/rel
Hunmorph	2005	0.3951	0.3900	883/939
lexicon+decomp	2006	0.3443	0.4320	1149/1308
Hunmorph	2005	0.3532	0.3560	836/939
lexicon	2006	0.2989	0.3840	1086/1308
Hunmorph	2005	0.3444	0.3400	819/939
no lexicon	2006	0.2785	0.3580	1025/1308
Snowball	2005	0.3371	0.3360	818/939
(no lexicon)	2006	0.2790	0.3820	965/1308
Savoy's	2005	0.3335	0.3360	819/939
(no lexicon)	2006	0.2792	0.3860	964/1308
baseline	2005	0.2153	0.2400	656/939
no stemmer	2006	0.1853	0.2700	757/1308

variants of Hunmorph also perform in the same range when it is used without a lexicon. In fact a comparison of different variants of this lexicon-free stemmer is instructive (see Table 3). Allowing for the recognition of derivational as well as inflectional suffixes (heavy) is always better than using only inflections (light). This is in line with or an extension to [7]’s finding that the more aggressive the stemming the better. It is noteworthy that stemming verbal suffixation on top of nominal can give worse results (as it does for 2005). It is uncertain whether this is because of an increased number of false noun-verb confluations. Pairwise comparisons between these variants and the two other stemmers show no significant difference.

Table 3. Comparison of different lexicon-free stemmers. nom-heavy: nominal suffixes, all-heavy: nominal and verbal suffixes; nom-light: nominal inflection only, all-light: nominal and verbal inflection only.

guesser	year	MAP	P10	ret/rel
nom-heavy	2005	0.3444	0.3400	819/939
	2006	0.2785	0.3580	1025/1308
all-heavy	2005	0.3385	0.3520	817/939
	2006	0.2837	0.3660	1019/1308
nom-light	2005	0.3244	0.3360	808/939
	2006	0.2653	0.3540	962/1308
all-light	2005	0.3158	0.3240	811/939
	2006	0.2663	0.3560	1000/1308

Note again, that Hunmorph uses a morphological grammar to guess alteernative stems. This has clearly beneficial consequences because its level of sophistication allows us to exclude erroneous suffix combinations that Snowball (maybe

mistakenly) identifies such as *alakot*→*al* instead of *alak* and which cause faulty conflations. On the other hand, Hunmorph allows all the various possibilities that may occur in the language with a lexicon however phonotactically implausible they are. The simple candidate selection heuristics does not take this into account and may choose (randomly) the implausible *ábrakat*→*ábrá* instead of *ábra* (the base form) and thereby missing crucial conflations.

One of the most important findings is that using a lexicon is not significantly better than any of the resourceless alternatives when compounding is not used.

As expected from the previous discussion, the best results are achieved when decomponing is used. Our method of decomponing ensures that component stems are (at least potentially) linguistically correct, i.e., composed of real existing stems with the right category matching a compound pattern. This also implies that only the head category (right constituent) can be inflected on a lexical data-base, though the other components can also be derived. Using this method of decomponing has a clearly beneficial effect on retrieval performance increasing accuracy from 34%/30% up to 39%/34% on the 2005 and the 2006 collection respectively (both years show that compounding is significantly better than all alternatives not using compounding). The 2005 result of 39% positions our system at the second place beating Savoy's system which uses the alternative compounding. A more direct comparison would require changing only the compounding algorithm while keeping the rest of our system unchanged. Since the reimplementaion of this system is non-trivial and there is no software available for the task, such a direct assessment of this effect is not available. Nonetheless, given the simplicity of our IR system, we suspect that our method of compounding performs significantly better. Systems in 2006 are usually much better, but this is primarily due to improvements in ranking and retrieval strategies. These improvements might in principle neutralize or make irrelevant any differences in compounding, however, the fact that most systems use decomponing allows us the conjecture that our advantage would carry over to more sophisticated IR systems.

We speculate that if some problems with guessing are mended, it might turn out to be more beneficial even when the full lexicon is used. As it stands guessing has only a non-significant positive effect.

Table 4 shows retrieval performance when POS tagger is used. We can see the unexpected result that contextual disambiguation has a negative effect on performance. Retrospectively, it is not too surprising that POS tagging brings no

Table 4. Result of different stemming methods *after* POS-tagging

	year	MAP	P10	ret/rel
without decomponing	2005	0.3289	0.3300	814/939
	2006	0.2855	0.3880	1009/1308
with decomponing	2005	0.3861	0.3740	893/939
	2006	0.3416	0.4360	1120/1308

improvements. First, for most of the tokens, there is nothing to disambiguate, either the analysis itself is unique or the lemmas belonging to the alternatives are identical. An example for the latter is so called syncretisms, situations where two different paradigmatic forms of a stem (class of stems) coincide. This is exemplified by the 1st person past indicative forms which are ambiguous between definite and indefinite conjugation.⁴ So the effort of the tagger to make a choice based on context is futile for retrieval. In fact, some of the hardest problems for POS tagging involve disambiguation of very frequent homonyms, such as *egy* ('one/NUM' or 'a/DET'), but most of these are discarded by stopword filtering anyway.

Using the POS tagger can even lead to new kinds of errors. For example, in Topic 367 of 2005, the analyzer gave two analyses for the word *drogok*: *drog*/PLUR inflected and the compound *drog+ok* ('drug-cause'); the POS tagger incorrectly chooses the latter. Such errors, though, could be fixed by blocking these arguably overgenerated compound analyses.

4 Conclusion

The experiments on which we report in this paper confirm that a lemmatization in Hungarian greatly improves retrieval accuracy. Our system outperformed all CLEF 2005 systems that use algorithmic stemmers despite its simplicity. Good results are due to the high-coverage lexical resources allowing decompounding (or lack thereof through blocking) and recognition of various derivational and inflectional patterns allowing for aggressive stemming.

We compared two different morphological analyzer-based lemmatization methods. We found that contextual disambiguation by a POS-tagger does not improve on simple local heuristics.

Our Hungarian lemmatizer (together with its morphological analyzer and a Hungarian descriptive grammar) is released under a permissive LGPL-style license, and can be freely downloaded from <http://mokk.bme.hu/resources/ir>. We hope that members of the CLEF community will incorporate these into their IR systems, closing the gap in effectivity between IR systems for Hungarian and for major European languages.

References

- [1] Halácsy, P.: Benefits of deep NLP-based lemmatization for information retrieval. In: Working Notes for the CLEF 2006 Workshop (September 2006)
- [2] McNamee, P.: Exploring New Languages with HAIRCUT at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 21–23. Springer, Heidelberg (2006)

⁴ Finite verbs in Hungarian agree with the object in definiteness. If the verb is intransitive or the object is an indefinite noun phrase, the indefinite conjugation has to be used.

- [3] Savoy, J., Berger, P.Y.: Report on CLEF-2005 Evaluation Campaign: Monolingual, Bilingual, and GIRT Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 21–23. Springer, Heidelberg (2006)
- [4] Goñi Menoyo, J.M., González, J.C., Vilena-Román, J.: Miracle’s 2005 approach to monolingual information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [5] Tomlinson, S.: European Ad hoc retrieval experiments with Hummingbird TM at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [6] Savoy, J.: Report on CLEF-2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, Springer, Heidelberg (2004)
- [7] Tordai, A., de Rijke, M.: Hungarian Monolingual Retrieval at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [8] Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of the ACL 2005 Workshop on Software (2005)
- [9] Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of LREC 2006, pp. 1670–1673 (2006)
- [10] Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Proceedings of the EACL 2006 Workshop on Web as a Corpus (2006)
- [11] Halácsy, P., Kornai, A., Oravecz, C., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC 2006, pp. 2245–2248 (2006)
- [12] Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of Language Resources and Evaluation Conference (LREC04), European Language Resources Association (2004)

Statistical vs. Rule-Based Stemming for Monolingual French Retrieval

Prasenjit Majumder¹, Mandar Mitra¹, and Kalyankumar Datta²

¹ CVPR Unit, Indian Statistical Institute, Kolkata

² Dept. of EE, Jadavpur University, Kolkata

{[@isical.ac.in](mailto:prasenjit_t,mandar),

kalyandatta@debesh.wb.nic.in

Abstract. This paper describes our approach to the 2006 Adhoc Monolingual Information Retrieval run for French. The goal of our experiment was to compare the performance of a proposed statistical stemmer with that of a rule-based stemmer, specifically the French version of Porter's stemmer. The statistical stemming approach is based on lexicon clustering, using a novel string distance measure. We submitted three official runs, besides a baseline run that uses no stemming. The results show that stemming significantly improves retrieval performance (as expected) by about 9-10%, and the performance of the statistical stemmer is comparable with that of the rule-based stemmer.

1 Introduction

We have recently been experimenting with languages that have not been studied much from the IR perspective. These languages are typically resource-poor, in the sense that few language resources or tools are available for them. As a specific example, no comprehensive stemming algorithms are available for these languages. The stemmers that are available for more widely studied languages (e.g. English) usually make use of an extensive set of linguistic rules. Rule based stemmers for most resource-poor languages are either unavailable or lack comprehensive coverage. In earlier work, therefore, we have looked at the problem of stemming for such resource-poor languages, and proposed a stemming approach that is based on purely unsupervised clustering techniques.

Since the proposed approach does not assume any language-specific information, we expect the approach to work for multiple languages. The motivation behind our experiments at CLEF 2006 was to test this hypothesis. Thus, we focused on mono-lingual retrieval for French (a language which we know nothing about), and tried our statistical stemming approach on French data.

We give a brief overview of the proposed statistical stemming algorithm in the next section. We outline our experimental setup in Section 3, and discuss the results of the runs that we submitted.

2 Statistical Stemmer

2.1 String Distance Measures

Distance functions map a pair of strings s and t to a real number r , where a smaller value of r indicates greater similarity between s and t . In the context of stemming, an appropriate distance measure would be one that assigns a low distance value to a pair of strings when they are morphologically similar, and assigns a high distance value to morphologically unrelated words. The languages that we have been experimenting with are primarily suffixing in nature, i.e. words are usually inflected by the addition of suffixes, and possible modifications to the tail-end of the word. Thus, for these languages, two strings are likely to be morphologically related if they share a long matching prefix. Based on this intuition, we define a string distance measure D which rewards long matching prefixes, and penalizes an early mismatch.

Given two strings $X = x_0x_1 \dots x_n$ and $Y = y_0y_1 \dots y_{n'}$, we first define a Boolean function p_i (for penalty) as follows:

$$p_i = \begin{cases} 0 & \text{if } x_i = y_i \quad 0 \leq i \leq \min(n, n') \\ 1 & \text{otherwise} \end{cases}$$

Thus, p_i is 1 if there is a mismatch in the i -th position of X and Y . If X and Y are of unequal length, we pad the shorter string with null characters to make the string lengths equal.

Let the length of the strings be $n + 1$, and let m denote the position of the first mismatch between X and Y (i.e. $x_0 = y_0, x_1 = y_1, \dots, x_{m-1} = y_{m-1}$, but $x_m \neq y_m$). We now define D as follows:

$$D(X, Y) = \frac{n - m + 1}{m} \times \sum_{i=m}^n \frac{1}{2^{i-m}} \quad \text{if } m > 0, \quad \infty \text{ otherwise} \quad (1)$$

Note that D does not consider any match once the first mismatch occurs. The actual distance is obtained by multiplying the total penalty by a factor which is intended to reward a long matching prefix, and penalize significant mismatches. For example, for the pair $\langle \text{astronomer}, \text{astronomically} \rangle$, $m = 8, n = 13$. Thus, $D_3 = \frac{6}{8} \times (\frac{1}{2^0} + \dots + \frac{1}{2^{13-8}}) = 1.4766$.

2.2 Lexicon Clustering

Using the distance function defined above, we can cluster all the words in a document collection into groups. Each group, consisting of “similar” strings, is expected to represent an equivalence class consisting of morphological variants of a single root word. The words within a cluster can be stemmed to the ‘central’ word in that cluster. Since the number of natural clusters are unknown apriori, partitive clustering algorithms like k -means are not suitable for our task. Also, the clusters are likely to be of non-convex nature. Graph-theoretic clustering

algorithms appear to be the natural choice in this situation because of their ability to detect natural and non-convex clusters in the data.

Three variants of graph theoretic clustering are popular in literature, namely, *single-linkage*, *average-linkage*, and *complete-linkage* [2]. Each of these algorithms are of hierarchical (agglomerative or divisive) nature. In the agglomerative form, the cluster tree (often referred to as a dendrogram) consists of individual data points as leaves. The nearest (or most similar) pair(s) of points are merged to form groups, which in turn are successively merged to form progressively larger groups of points. Clustering stops when the similarity between the pair of closest groups falls below a pre-determined threshold. Alternatively, a threshold can be set on the distance value; when the distance between the pair of nearest points exceeds the threshold, clustering stops. The three algorithms mentioned above differ in the way similarity between the groups is defined. We choose the complete-linkage algorithm for our experiments.

3 Results

We used the Smart [3] system for all our experiments. We submitted four official runs, including one baseline. For the baseline run (Cbaseline), queries and documents were indexed after eliminating stopwords (using the stopwords list provided on the CLEF website¹). The <title>, <desc>, and <narr> field of the query were indexed. The *Lnu.ltn* term-weighting strategy [1] was used. No stemming was done for the baseline run.

For the remaining three runs, we used three variants of the statistical stemming method described above. Since our approach is based on hierarchical agglomerative clustering (as described above), the threshold value used in the clustering step is an important parameter of the method. Earlier experiments with English data have shown that 1.5 is a reasonable threshold value. We generated two retrieval runs by setting the threshold to 1.5 and 2.0 respectively (Cd61.5, Cd62.0).

For the third run, the stemmer was created based on a subset of the data. A lexicon was constructed using only the LeMonde section of the document collection, and this was then clustered as described above to determine the stem classes. Since the lexicon was smaller, the clustering step took less time for this run. The motivation behind this experiment was to study how performance is affected when a subset of the lexicon is used to construct the stemmer in order to save computation time.

After the relevance judgments for this data set were distributed, we performed two additional experiments: first, we tried setting the clustering threshold to 1.0; and secondly, we used the French version of Porter's stemmer² in place of our statistical stemmer. The results obtained for all the official and unofficial runs are given below.

¹ <http://www.unine.ch/info/clef/>

² Downloaded from <http://www.snowball.tartarus.org/algorithms/french/stemmer.html>

Table 1. Retrieval results obtained using various stemmers

Run ID	Topic fields	MAP	Rel. ret.	R-Precision
Cbaseline	T+D+N	0.3196	1,616	31.30%
Cd61.0	T+D+N	0.3465 (+8.4%)	1,715	34.85%
Cd61.5 (official)	T+D+N	0.3454 (+8.1%)	1,709	34.35%
Cd61.5 (obtained)	T+D+N	0.3509 (+9.8%)	1,708	34.26%
Cd62.0	T+D+N	0.3440 (+7.6%)	1,737	34.78%
Cld61.5	T+D+N	0.3342 (+4.6%)	1,678	32.81%
Porter	T+D+N	0.3480 (+8.9%)	1,705	34.71%

The official results for the run labelled Cd61.5 do not agree with the evaluation figures that we obtained by using the distributed relevance judgment data. We therefore report both the official figures and the numbers that we obtained in Table 1.

The two most promising runs (Cd61.5 and Porter) were analyzed in greater detail. Paired *t*-tests show that stemming (using either strategy) results in significant improvements (at a 1% level of confidence) over the baseline (no stemming), but the differences between the rule-based and statistical approaches are not statistically significant. Also, some loss in performance results when the stemmer is generated from a subset of the corpus (run Cld61.5).

This confirms our hypothesis that the proposed stemming approach, which does not assume any language-specific information, will work for a variety of languages, provided the languages are primarily suffixing in nature.

References

1. Buckley, C., Singhal, A., Mitra, M.: Using Query Zoning and Correlation within SMART: TREC5. In: Voorhees, E.M., Harman, D.K. (eds.) Proceedings of the Fifth Text REtrieval Conference (TREC-5). NIST Special Publication, pp. 500–238 (November 1997)
2. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. *ACM Comput. Surv.* 31(3), 264–323 (1999)
3. Salton, G.: The SMART Retrieval System—Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs (1971)

A First Approach to CLIR Using Character N -Grams Alignment

Jesús Vilares¹, Michael P. Oakes², and John I. Tait²

¹ Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 - A Coruña Spain
jvilares@udc.es

² School of Computing and Technology, University of Sunderland
St. Peter's Campus, St. Peter's Way, Sunderland - SR6 0DD United Kingdom
{Michael.Oakes, John.Tait}@sunderland.ac.uk

Abstract. This paper describes the technique for translation of character n -grams we developed for our participation in CLEF 2006. This solution avoids the need for word normalization during indexing or translation, and it can also deal with out-of-vocabulary words. Since it does not rely on language-specific processing, it can be applied to very different languages, even when linguistic information and resources are scarce or unavailable. Our proposal makes considerable use of freely available resources and also tries to achieve a higher speed during the n -gram alignment process with respect to other similar approaches.

Keywords: Cross-Language Information Retrieval, character n -grams, translation algorithms, alignment algorithms, association measures.

1 Introduction

This work has been inspired by the previous approach of the Johns Hopkins University Applied Physics Lab (JHU/APL) on the employment of overlapping character n -grams for indexing documents [7,8]. Their interest came from the possibilities that overlapping character n -grams may offer particularly in the case of non-English languages: to provide a surrogate means to normalize word forms and to allow to manage languages of very different natures without further processing. This knowledge-light approach does not rely on language-specific processing, and it can be used even when linguistic information and resources are scarce or unavailable.

In the case of monolingual retrieval, the employment of n -grams is quite straight forward, since both queries and documents are just tokenized into overlapping n -grams instead of words. In the case of cross-language retrieval, two phases are required during query processing: translation and n -gram splitting. In their later experiments, JHU/APL developed a new *direct n -gram translation* technique that used n -grams instead of words as translation units. Their goal was to avoid some of the limitations of classical dictionary-based translation, such as the need for word normalization or the inability to handle out-of-vocabulary words. This n -gram translation algorithm takes as input a parallel

corpus, aligned at the paragraph (or document) level and extracts candidate translations as follows [8]. Firstly, for each candidate n -gram term to be translated, paragraphs containing this term in the source language are identified. Next, their corresponding paragraphs in the target language are also identified and, using a statistical measure similar to mutual information, a translation score is calculated for each of the terms occurring in one of the target language texts. Finally, the target n -gram with the highest translation score is selected as the potential translation of the source n -gram. The whole process is quite slow: it is said that the process takes several days in the case of working with 5-grams, for example.

This paper describes our first experiments in the field of Cross-Language Information Retrieval (CLIR) employing our own direct n -gram translation technique. This new approach also tries to achieve a higher speed during the n -gram alignment process in order to make easier the testing of new statistical measures. The article is structured as follows. Firstly, Sect. 2 describes our n -gram-based CLIR system. Next, Sect. 3 shows the results of our first tests, still in development. Finally, Sect. 4 presents our preliminary conclusions and future work.

2 The Character N -Gram Alignment Algorithm

Taking as our model the system designed by JHU/APL, we developed our own n -gram based retrieval system. Instead of the ad-hoc resources developed for the original system [7,8], our system has been built using freely available resources when possible in order to make it more transparent and to minimize effort.

This way, we have opted for using the open-source retrieval platform TERIER [1]. This decision was supported by the satisfactory results obtained with n -grams using different indexing engines [10].

A second difference comes from the translation resources to be used, in our case the well-known EUROPARL parallel corpus [4]. This corpus was extracted from the proceedings of the European Parliament covering April 1996 to September 2003, containing up to 28 million words per language. It includes versions in 11 European languages: Romance (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

Finally, with respect to the n -gram translation algorithm itself, it now consists of two phases. In the first phase, the slowest one, the input parallel corpus is aligned at word-level using the well-known statistical aligner GIZA++ [9], obtaining the translation probabilities between the different source and target language words. Next, prior to the second phase, several heuristics can be applied —if desired— for refining or modifying such word-to-word translation scores. We can remove, for example, the least probable candidate translations, or combine the scores of bidirectional alignments [5] —source-target language and target-source language— instead of using just the direct one —source-target language. Finally, in the second phase, n -gram translation scores are computed employing statistical association measures [6], taking as input the translation probabilities calculated by GIZA++.

This approach increases the speed of the process by concentrating most of the complexity in the word-level alignment phase. This first step acts as a initial filter, since only those n -gram pairs corresponding to aligned words will be considered, whereas in the original JHU/APL approach all n -gram pairs corresponding to aligned paragraphs were considered. On the other hand, since the n -gram alignment phase is much faster, different n -gram alignment techniques can be easily tested. Another advantage of this approach is that the n -gram alignment process can take as input previously existing lists of aligned words or even bilingual dictionaries, theoretically improving the results.

2.1 Word-Level Alignment Using Association Measures

In order to better illustrate the process involved in this second phase, we will take as basis how association measures could be used for creating bilingual dictionaries taking as input parallel collections aligned at paragraph level. In this context, given a word pair $(word_u, word_v)$ — $word_u$ standing for the source language word, and $word_v$ for its candidate target language translation—, their cooccurrence frequency can be organized in a *contingency table* resulting from a cross-classification of their cooccurrences in the aligned corpus:

$ V = word_v V \neq word_v $			
$U = word_u$	O_{11}	O_{12}	$= R_1$
$U \neq word_u$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

In this table, instances whose first component belongs to type $word_u$ —i.e., the number of aligned paragraphs where the source language paragraph contains $word_u$ — are assigned to the first row of the table, and tokens whose second component belongs to type $word_v$ —i.e., the number of aligned paragraphs where the target language paragraph contains $word_v$ — are assigned to the first column. The cell counts are called the *observed frequencies*: O_{11} , for example, stands for the number of aligned paragraphs where the source language paragraph contains $word_u$ and the target language paragraph contains $word_v$. The sum of the observed frequencies —or *sample size* N — is the total number of word pairs considered. The row totals, R_1 and R_2 , and the column totals, C_1 and C_2 , are also called *marginal frequencies*, and O_{11} is called the *joint frequency*.

Once the contingency table has been built, different association measures can be easily calculated for each word pair. The most promising pairs, those with the highest association measures, will take part of the bilingual dictionary. Our system employs two classical measures: *mutual information* and the *Dice coefficient*, defined by equations 1 and 2, respectively:

$$MI(word_u, word_v) = \log \frac{NO_{11}}{R_1 C_1} \quad (1) \quad Dice(word_u, word_v) = \frac{2O_{11}}{R_1 + C_1} \quad (2)$$

2.2 Adaptations for *N*-Gram-Level Alignment

We have described how to compute and employ association measures for generating bilingual dictionaries from parallel corpora aligned at the paragraph level. However, in our proposal, we do not have aligned paragraphs but aligned words—a source word and its candidate translation—, both composed of *n*-grams. Our first idea could be just to adapt the contingency table to this context, by considering that we are now dealing with *n*-gram pairs (*n-gram_u*, *n-gram_v*) cooccurring in aligned words instead of word pairs (*word_u*, *word_v*) cooccurring in aligned paragraphs. So, contingency tables should be redefined according to this new situation: *O*₁₁, for example, should be re-formulated as the number of aligned word pairs where the source language word contains *n-gram_u* and the target language word contains *n-gram_v*.

This solution seems logical, but there is a problem. In the case of aligned paragraphs, we had *real* instances of word cooccurrences at the paragraphs aligned. However, now we do not have *real* instances of *n*-gram cooccurrences at aligned words—as may be expected—, but just *probable* ones, since GIZA++ uses a statistical alignment model which computes a translation probability for each cooccurring word pair. So, the same word may appear as being aligned with several translation candidates, each one with a given probability. For example, taking the English words *milk* and *milky*, and the Spanish words *leche* (*milk*), *lechoso* (*milky*) and *tomate* (*tomato*), a possible output alignment would be:

source word	candidate translation	probability
milk	leche	0.98
milky	lechoso	0.92
milk	tomate	0.15

This way, it may be considered that the source 4-gram *-milk-* does not really cooccur with the target 4-gram *-lech-*, since the alignment between its containing words *milk* and *leche*, and *milky* and *lechoso* is not certain. Nevertheless, it seems much more probable that the *translation* of *-milk-* was *-lech-* rather than *-toma-*, since the probability of the alignment of their containing words—*milk* and *tomato*—is much smaller than that of the words containing *-milk-* and *-lech-*—the pairs *milk* and *leche* and *milky* and *lechoso*. Taking this idea as a basis, our proposal consists of weighting the likelihood of a cooccurrence according to the probability of its corresponding alignment. So, the contingency tables corresponding to the *n*-gram pairs (*-milk-*, *-lech-*) and (*-milk-*, *-toma-*) are as follows:

	<i>V</i> = <i>-lech-</i>	<i>V</i> ≠ <i>-lech-</i>	
<i>U</i> = <i>-milk-</i>	$ O_{11} = 0.98 + 0.92 = \mathbf{1.90} $	$ O_{12} = 0.98 + 3 * 0.92 + 3 * 0.15 = \mathbf{4.19} $	$ R_1 = \mathbf{6.09} $
<i>U</i> ≠ <i>-milk-</i>	$ O_{21} = \mathbf{0.92} $	$ O_{22} = 3 * 0.92 = \mathbf{2.76} $	$ R_2 = \mathbf{3.68} $
	$ C_1 = \mathbf{2.82} $	$ C_2 = \mathbf{6.95} $	$ N = \mathbf{9.77} $

$ V = \text{-toma-} $	$V \neq \text{-toma-}$		
$U = \text{-milk-} $	$O_{11} = \mathbf{0.15}$	$ O_{12} = 2 * 0.98 + 4 * 0.92 + 2 * 0.15 = \mathbf{5.94} $	$R_1 = \mathbf{6.09}$
$U \neq \text{-milk-} $	$O_{21} = \mathbf{0}$	$O_{22} = 4 * 0.92 = \mathbf{3.68}$	$R_2 = \mathbf{3.68}$
$ C_1 = \mathbf{0.15} $	$C_2 = \mathbf{9.62}$		$ N = \mathbf{9.77}$

It can be seen that the O_{11} frequency corresponding to the n -gram pair (*-milk-*, *-lech-*) is not 2 as might be expected, but 1.90. This is because it appears in 2 alignments, *milk* with *leche* and *milky* with *lechoso*, but each cooccurrence in a alignment must also be weighted according to its translation probability like this: 0.98 (for *milk* with *leche*) + 0.92 (for *milky* with *lechoso*) = 1.90 .

Once the contingency tables have been obtained, the Dice coefficients corresponding to each n -gram pair can be computed, for example. As expected, the association measure of the pair (*-milk-*, *-lech-*) —the correct one— is much higher than that of the pair (*-milk-*, *-toma-*) —the wrong one:

$$Dice(\text{-milk-}, \text{-lech-}) = \frac{2 * 1.90}{6.09 + 2.82} = \mathbf{0.43} \quad Dice(\text{-milk-}, \text{-toma-}) = \frac{2 * 0.15}{6.09 + 0.15} = \mathbf{0.05}$$

3 Evaluation

Our group took part in the CLEF 2006 ad-hoc track [11], but the lack of time did not allow us to complete our n -gram direct translation tool. So, we could submit only those results intended to be used as baselines for future tests. Since these results are publicly available in [2], we will discuss here only those new results obtained in later experiments.

These new experiments were made with our n -gram direct translation approach using the English topics and the Spanish document collections of the robust task —i.e., a English-to-Spanish run. The *robust task* is essentially an ad-hoc task which takes the topics and collections used from CLEF 2001 to CLEF 2003. In the case of the Spanish data collection, it is formed by 454,045 news reports (1.06 GB) provided by EFE, a Spanish news agency, corresponding to the years 1994 and 1995. The test set consists of 160 topics (C041–C200). This initial set is divided into two subsets: a *training topics* subset to be used for tuning purposes and formed by 60 topics (C050–C059, C070–C079, C100–C109, C120–C129, C150–159, C180–189), and a *test topics* subset for testing purposes. Since the goal of these first experiments is the tuning and better understanding of the behavior of the system, we will only use here the *training topics* subset. Moreover, only *title* and *description* fields were used in the submitted queries.

With respect to the indexing process, documents were simply split into n -grams and indexed, as were the queries. We use 4-grams as a compromise n -gram size after studying the results previously obtained by the JHU/APL group [7,8] using different lengths. Before that, the text had been lowercased and punctuation marks were removed [8], but not diacritics. The open-source TERRIER platform [1] was used as retrieval engine with a InL2¹ ranking model [3]. No stopwords removal or query expansion have been applied at this point.

¹ Inverse Document Frequency model with Laplace after-effect and normalization 2.

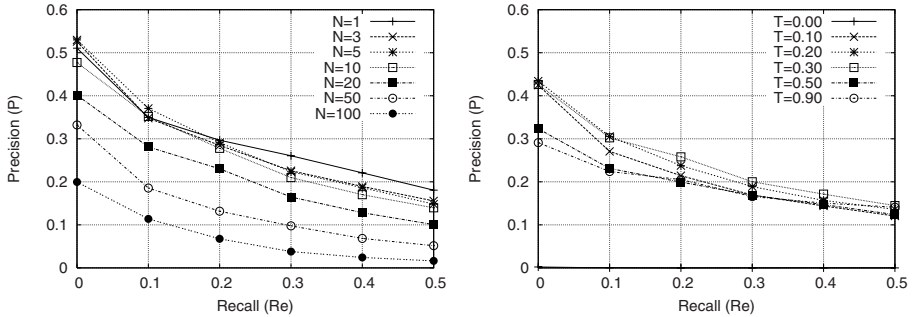


Fig. 1. Precision vs. Recall graphs when taking the N most probable n -gram translations (*left*) and when using a minimal probability threshold T (*right*)

For querying, the source language topic is firstly split into n -grams, 4-grams in our case. Next, these n -grams are replaced by their candidate translations according to a selection algorithm. Two algorithms are available: the first one takes the N most probable alignments, and the second one takes those alignments with a probability greater or equal than a threshold T . The resulting translated topics are then submitted to the retrieval system.

With respect to the n -gram alignment itself, we have refined the initial word-level alignment by using bidirectional alignment. That is, we will consider a $(word_{English}, word_{Spanish})$ English-to-Spanish alignment only if there is a corresponding $(word_{Spanish}, word_{English})$ Spanish-to-English alignment. Finally, we have used only one of the association measures available, the Dice coefficient.

The results for this first approach are shown in the two Precision vs. Recall graphs of Fig. 1². The figure on the left shows the results obtained when taking the N most probable n -gram translations, with $N \in \{1, 3, 5, 10, 20, 50, 100\}$. The figure on the right shows the results obtained when using a minimal probability threshold T , with $T \in \{0.00, 0.10, 0.20, 0.30, 0.50, 0.90\}$. As it can be seen, the results taking the N most probable alignments are better, particularly when using few translations.

Next, trying to improve the accuracy of the n -gram alignment process, we removed those least-probable word alignments from the input (those with a word translation probability less than a threshold W , with $W=0.15$). The new results obtained, shown in Fig. 2, are very similar to those without pruning. Nevertheless, such pruning led to a considerable reduction of processing time and storage space: 95% reduction in the number of input word pairs processed and 91% reduction in the number of output n -gram pairs aligned. The results taking the N most probable alignments improve as N is reduced.

Finally, Fig. 3 shows our best results with respect to several baselines: by querying the Spanish index with the English topics split into 4-grams (EN_{4gr}) —for measuring the impact of casual matches—, by using stemmed Spanish topics (ES_{stm}), and by using Spanish topics split into 4-grams (ES_{4gr}) —our

² Highest recall levels, lesser in importance, have been removed for improving reading.

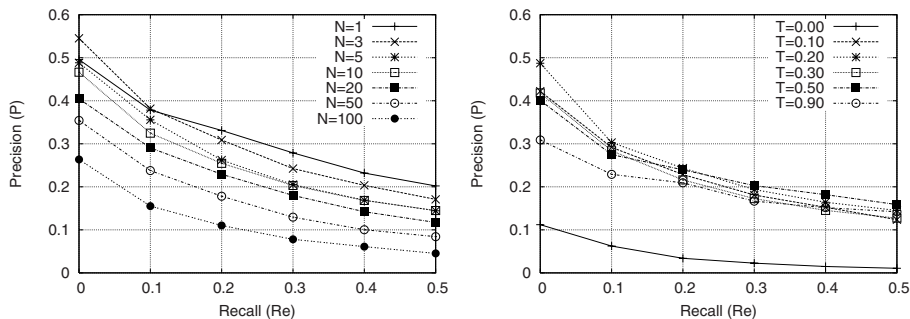


Fig. 2. Precision vs. Recall graphs when taking the N most probable n -gram translations (*left*) and when using a minimal probability threshold T (*right*). Input word alignments with a probability lesser than $W=0.15$ have been dismissed.

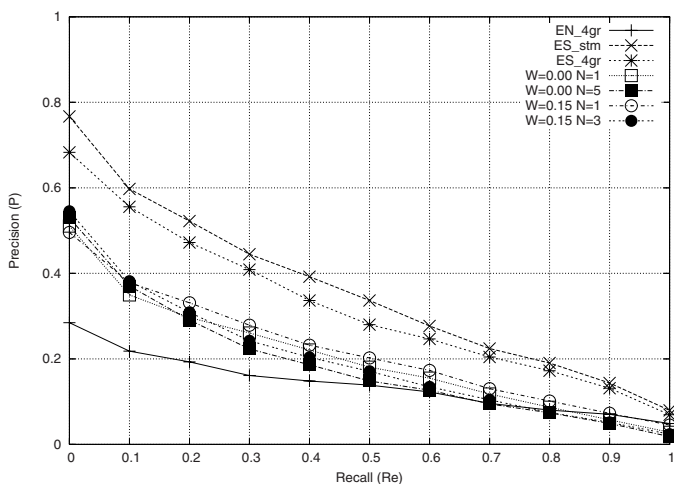


Fig. 3. Summary Precision vs. Recall graph

performance goal. Although current performance is not as good as expected, these results are encouraging, since it must be taken into account that these are our very first experiments, so the margin for improvement is still great.

4 Conclusions and Future Work

This paper describes our initial work in the field of Cross-Language Information Retrieval in developing a system which uses character n -grams not only as indexing units, but also as translation units. This system was inspired by the work of the Johns Hopkins University Applied Physics Lab [7,8], but freely available resources were used when possible. Our n -gram alignment algorithm consists of two phases. In the first phase, the slowest one, word-level alignment of the text

is made through a statistical alignment tool. In the second phase, n -gram translation scores are computed employing statistical association measures, taking as input the translation probabilities calculated in the previous phase. This new approach speeds up the training process, concentrating most of the complexity in the word-level alignment phase, making the testing of new association measures for n -gram alignment easier.

With respect to future work, once tuned, the system will be tested with other bilingual and multilingual runs. The employment of relevance feedback, and the use of pre or post-translation expansion techniques, are also planned. Finally, we also intend to try new association measures [12] for n -gram alignment.

Acknowledgment

This research has been partially supported by Ministerio de Educación y Ciencia and FEDER (TIN2004-07246-C03-02), Xunta de Galicia (PGIDIT05PXIC30501PN, PGIDIT05SIN044E), and Dirección Xeral de Investigación, Desenvolvemento e Innovación (*Programa de Recursos Humanos* grants).

References

1. <http://ir.dcs.gla.ac.uk/terrier/>
2. <http://www.clef-campaign.org>
3. Amati, G., van Rijsbergen, C.J.: Probabilistic models of Information Retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems* 20(4), 357–389 (2002)
4. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proc. of the 10th Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 12–16, 2005, pp. 79–86 (2005), Corpus available in <http://www.iccs.inf.ed.ac.uk/~pkoeht/publications/europarl/>
5. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the ACL*, Morristown, NJ, USA, pp. 48–54 (2003)
6. Manning, C.D., Schütze, H.: *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge (1999)
7. McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7(1–2), 73–97 (2004)
8. McNamee, P., Mayfield, J.: JHU/APL experiments in tokenization and non-word translation. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) *CLEF 2003*. LNCS, vol. 3237, pp. 85–97. Springer, Heidelberg (2004)
9. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models, *Toolkit* (2003), available at <http://www.fjoch.com/GIZA++.html>
10. Savoy, J.: Cross-Language Information Retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management* 39, 75–115 (2003)
11. Vilares, J., Oakes, M.P., Tait, J.I.: CoLesIR at CLEF 2006: rapid prototyping of a N-gram-based CLIR system. In: *Working Notes of the CLEF 2006 Workshop*, 20–22 September, Alicante, Spain (2006) available at [2]
12. Weeds, J., Weir, D.: Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics* 31(4), 439–475 (2005)

SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: Query Expansion Using the Google Search Engine

Fernando Martínez-Santiago, Arturo Montejo-Ráez,
Miguel Á. García-Cumbreras, and L. Alfonso Ureña-López

SINAI Research Group. Computer Science Department. University of Jaén. Spain
{dofer, amontejo, magc, laurena}@ujaen.es

Abstract. This year, we have participated in the Ad-Hoc Robust Multilingual track with the aim of evaluating two important issues in Cross-Lingual Information Retrieval (CLIR) systems. This paper first describes the method applied for query expansion in a multilingual environment by using web search results provided by the Google engine in order to increase retrieval robustness. Unfortunately, the results obtained are disappointing. The second issue reported alludes to the robustness of several common merging algorithms. We have found that 2-step RSV merging algorithms perform better than others algorithms when evaluating using geometric average¹.

1 Introduction

Robust retrieval has been a task in the TREC evaluation forum [1]. One of the best systems proposed involves query expansion through web assistance [4,3,2]. We have followed the approach of Kwok and his colleagues and applied it to robust multilingual retrieval.

Pseudo-relevance feedback has been used traditionally to generate new queries from the results obtained from a given source query. Thus, the search is launched twice: one for obtaining first relevant documents from which new query terms are extracted, and a second turn to obtain final retrieval results. This method has been found useful to solve queries producing small result sets, and is a way to expand queries with new terms that can widen the scope of the search. But pseudo-relevance feedback is not that useful when queries are so difficult that very few or no documents are obtained at a first stage (the so-called weak queries). In that case, there is a straightforward solution: use a different and richer collection to expand the query. Here, the Internet plays a central role: it is a huge amount of web pages where virtually any query, no matter how difficult it is, may be related to some subset of those pages. This approach has obtained

¹ This work has been supported by the Spanish Government (MCYT) with grant TIC2003-07158-C04-04 and the RFC/PP2006/Id_514 granted by the University of Jaén.

remarkable results in monolingual IR systems evaluated at TREC conferences. Unexpectedly, the results obtained in a multilingual setting are very poor and we think that our implementation of the approach must be tuned for CLEF queries, regardless of our conviction that intensive tuning is unrealistic for real-world systems. In addition, as we expected, the quality of the expanded terms depends on the selected language.

In addition, we have evaluated several merging algorithms from the perspective of robustness: round-Robin, raw scoring, normalized raw scoring, logistic regression, raw mixed 2-step RSV, mixed 2-step RSV based on logistic regression and mixed 2-step RSV based on bayesian logistic regression. We have found that round-Robin, raw scoring and methods based on logistic regression perform worse than 2-step RSV merging algorithms.

The rest of the paper has been organized as three main sections: first, we describe the experimentation framework, then we report our bilingual experiments with web-based expansion queries, and finally we describe the multilingual experiments and how geometric precision affects several merging algorithms.

2 Experimentation Framework

Our Multilingual Information Retrieval System uses English as the selected topic language. The goal is to retrieve relevant documents for all languages in the collection, listing the results in a single, ranked list. In this list there is a set of documents written in different languages retrieved as an answer to a query in a given language (here, English). There are several approaches to this task, such as translating the whole document collection to an intermediate language or translating the question to every language found in the collection. Our approach is the latter: we translate the query for each language present in the multilingual collection. Thus, every monolingual collection must be preprocessed and indexed separately, as is described below.

2.1 Preprocessing and Translation Resources

In CLEF 2006 the multilingual task is made up of six languages: Dutch, English, French, German, Italian and Spanish. The pre-processing of the collections is the usual language specific in CLIR, taking into account the lexicographical and morphological idiosyncrasy of every language. The pre-processing is summarized in table 1.

Table 1. Language preprocessing and translation approach

	Dutch	English	French	German	Spanish	Italian
Preprocessing	stop words removed and stemming					
Decompounding	yes	no	no	yes	no	yes
Translation approach	FreeTrans		Reverso	Reverso	Reverso	FreeTrans

- English has been pre-processed as in our participation in the last years. Stop-words have been eliminated and we have used the Porter algorithm[7] as implemented in the ZPrise system.
- Dutch, German and Swedish are agglutinating languages. Thus, we have used the compounding algorithm depicted in [6]. The stopwords list and the stemmer algorithm have been both obtained in the Snowball site ².
- The resources for French and Spanish have been updated using the stop-word lists and stemmers from <http://www.unine.ch/info/clef>. The translation from English has been done using Reverso³ software.
- Dutch and Swedish translations have been done using on-line FreeTrans service⁴.

Once the collections have been pre-processed, they are indexed with the IR-N [10] system, an IR system based on passage retrieval. The OKAPI model has also been used for the on-line re-indexing process required by the calculation of 2-step RSV, using the OKAPI probabilistic model (fixed empirically at $b = 0.75$ and $k1 = 1.2$) [8]. As usual, we have not used blind feedback, because the improvement is marginal for these collections, and the precision is even worse for some languages (English and Swedish).

2.2 Merging Strategies

This year, we have selected the following merging algorithms: round-Robin, raw scoring [11,14], normalized raw scoring [13], logistic regression [12], raw mixed 2-step RSV, mixed 2-step RSV based on logistic regression [6] and mixed 2-step RSV based on bayesian logistic regression as implemented in the BBR package⁵.

3 Query Expansion Using the Internet as a Resource

Expanding user queries by using web search engines such as Google has been successfully used for improving the robustness of retrieval systems over collections in English. Due to the multilinguality of the web, we have assumed that this could be extended to additional languages, though the smaller amount of web pages not in English could be a major obstacle. The process is splitted into the following sequential steps:

1. **Web query generation.** The process varies depending on whether we consider the title field or the description field of the original query:
 - *From title.* Experiments expanding queries based just on the title field take all the terms in the field in lower case joined by the AND operator.

² Snowball is a small string-handling language in which stemming algorithms can be easily represented. Its name was chosen as a tribute to SNOBOL and is available on-line at <http://www.snowball.tartarus.org>

³ Reverso is available on-line at www.reverso.net

⁴ FreeTrans is available on-line at www.freetranslation.com

⁵ BBR software available at <http://www.stat.rutgers.edu/~madigan/BBR>.

- *From description.* Here, terms have to be selected. To that end, stop words are removed (using a different list according to the language which the description is written in) and the top five ranked terms are taken to compose, as for the title field, an AND query. The score is computed for each term to rank the same formula used by Kwok [3].
- 2. **Web search.** Once the web query has been composed, the web search engine is called to retrieve relevant documents. We can automate the process of query expansion through Google using its Java API. This web search is done specifying the language of the documents expected in the result set. Therefore, a filter on the language is set on the Google engine.
- 3. **Term selection from web results.** The 20 top ranked items returned by Google are taken. Each item points at an URL but also contains the so-called snippet, which is a selection of text fragments from the original pages containing the terms used in the web search (i.e. the query terms) as a sort of summary intended to let the user decide if the link is worth following. We have performed experiments using just the snippets as retrieved text in order to propose new query terms, and also experiments where terms are selected from full web page content (downloading the document from the returned URL).

In both cases (selection of terms from snippets or selection from full web pages), the final set of terms is the composite of those 60 terms with the highest frequency after discarding stop words. Of course, in the case of full web pages, the HTML tags are also conveniently eliminated. To generate the final expanded query, terms are added according to their frequencies (normalized to that of the least frequent term in the group of 60 selected terms).

As an example of the queries generated by this process, for a title containing words “inondation pays bas allemagne” the resulting expansion would produce the text:

```
pays pays pays pays pays pays pays pays pays pays pays pays pays pays
pays pays pays bas bas bas bas bas bas bas bas bas bas bas bas bas
allemagne allemagne allemagne allemagne allemagne allemagne allemagne
inondations inondations inondations france france france inondation
inondation inondation sud sud cles cles belgique belgique grandes
grandes histoire middot montagne delta savent fluviales visiteurs
exportateur engag morts pend rares projet quart amont voisins ouest
suite originaires huiti royaume velopp protection luxembourg convaincues
galement taient dues domination franque xiii tre rent commenc temp
monarchie xii maritime xive proviennent date xiiie klaas xiie ques
```

3.1 Experiments and Results

We have generated four different sets of queries for every language, one without expansion and three with web-based expansion:

1. **base** – No expansion, the original query is used and its results taken as base case
2. **sd-esnp** – Expansion using the original description field for web query generation and final terms selected from snippets
3. **st-esnp** – Expansion using the original title field for web query generation and final terms selected from snippets
4. **st-efpg** – Expansion using the original title field for web query generation and final terms selected from full web pages

The results obtained are discouraging, as all our expansions lead to worse measurements of both *R-precision* and *average precision*. Figures 1 a and 1 b show graphically the values obtained when evaluating these measures. For technical reasons, the expansion of type **st-efpg** for Dutch was not generated.

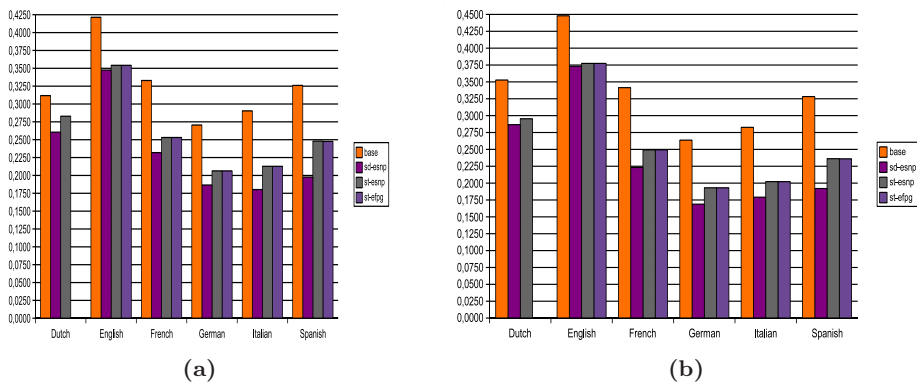


Fig. 1. R-precision (a) and average precision (b) measurements

Some conclusions can be drawn from these results. The main one is that the title field is a much more suitable source of items for a web-based expansion. Indeed, for many authors the title can be considered as the set of query terms that the users should pass to a search engine. Thus, web query generation from the description field, even using sophisticated formulae, is, as the results show, a worse choice when a title field is available.

We can check that very similar results are obtained, independently of the final selection of terms, that is, it seems that the decision of taking final terms either from snippets or from full web pages text does not determine significant differences in the results obtained. This issue needs further analysis, because expanded queries are quite different on the last half of the series of selected terms (those that are less frequent). These results hint at the system not profiting from the full set of terms passed, though possible errors on the experimental system may not have been identified yet.

Finally, we find that the results depend on the language under study. We think that this is due to differences on the size of existing collections of pages for each language on the web. That could explain the slightly better results in the case of English compared to the rest of languages.

In any case, these results demand for a deeper analysis in which different parameters are placed under study. For instance, the use of the AND operator in Google queries force to get only those pages with all the terms in the query, which is not desirable when expansion leads to large sets of query terms.

4 Multilingual Experiments

As the commented formerly, the merging algorithm is the only difference between all our multilingual experiments. Table 2 shows the results obtained in terms of 11-pt average precision, R-precision and geometric precision. From the point of view of the average precision, the most interesting finding is the relatively poor result obtained with the methods based on machine learning. Thus, mixed 2-step RSV-LR (Logistic Regression) and mixed 2-step RSV-BLR (Bayesian Logistic Regression) perform slightly worse than raw mixed 2-step RSV in spite that the latter approach does not use any training data. As usual, logistic regression performs better than round-Robin and raw scoring, but the difference is not as relevant as other years. Thus, we think that difficult queries are not learned as fine as usual queries. This is probably because, given a hard query, the relation between score, ranking and relevance of a document is not clear at all, and therefore, machine learning approaches are not capable to learn a good enough prediction function. Similarly, this year there are not only hard queries, but also very heterogeneous queries from the point of view of average precision. Thus, the distribution of average precision is very smooth and it makes it more difficult to extract useful information from the training data.

Table 2. Multilingual results

Merging approach	11Pt-AvgP	R-precision	Geometric Precision
round-robin	23.20	25.21	10.12
raw scoring	22.12	24.67	10.01
normalized Raw scoring	22.84	23.52	10.52
logistic regression	25.07	27.43	12.32
raw mixed 2-step RSV	27.84	32.70	15.70
mixed 2-step RSV based on LR	26.91	30.50	15.13
mixed 2-step RSV based on BLR	26.04	31.05	14.93

Since the 2-step RSV largely overcomes the rest of the tested merging algorithms when they are evaluated using geometric precision measure, we think that a 2-step RSV merging algorithm is better suited than other merging algorithms to improve the robustness of CLIR systems. Thus, if we use geometric

precision to evaluate the CLIR system, the difference of performance between results obtained using 2-step RSV and the rest of merging algorithms is higher than using traditional 11Pt-AvP or R-precision measures.

5 Conclusions

We have reported on our experimentation for the Ad-Hoc Robust Multilingual track CLEF task about web-based query expansion for languages other than English. First, we try to apply the expansion of queries using a web search engine such as Google. This approach has obtained remarkable results in monolingual IR systems evaluated at TREC conferences. But in a multilingual scenario, the results obtained are very poor and we think that our implementation of the approach must be tuned for CLEF queries, regardless of our belief that intensive tuning is unrealistic for real-world systems. In a robust evaluation the key measure should be the *geometric average precision*, also at monolingual tasks, because it emphasizes the effect of improving for weak queries, as the task itself defines. Future research includes to study the value obtained on this measure when using expanded queries and when merging retrieved items in multilingual retrieval, as it is difficult to explain the goodness of our approach to the robust task without it.

In addition, as was expected, the quality of the expanded terms depends on the language selected. The degradation in performance could have been produced by noisy expansions. As our expansions are considered totally new queries and are evaluated apart from original ones, we may discard a potential synergy between two resulting ranked lists. In this sense, we plan to merge both lists and evaluate it in terms of robustness.

The second issue reported is about the robustness of several widespread merging algorithms. We have found that Round-Robin, raw scoring and methods based on logistic regression perform worst from the point of view of robustness. In addition, 2-step RSV merging algorithms perform better than the other algorithms when geometric precision is applied. In any case, we think that the development of a robust CLIR system does not require special merging approaches: it "only" requires good merging approaches. It may be that other CLIR problems such as translation strategies or the development of an effective multilingual query expansion should be revisited in order to obtain such a robust CLIR model.

References

1. Voorhees, E.M.: The TREC Robust Retrieval Track, TREC Report (2005)
2. Kwok, K.L., Grunfeld, L., Deng, P.: Improving Weak Ad-Hoc Retrieval by Web Assistance and Data Fusion. In: Lee, G.G., Yamada, A., Meng, H., Myaeng, S.-H. (eds.) AIRS 2005. LNCS, vol. 3689, pp. 17–30. Springer, Heidelberg (2005)
3. Kwok, K.L., Grunfeld, L., Sun, H.L., Deng, P.: TREC2004 Robust Track Experiments using PIRCS, 2004 (2005)

4. Grunfeld, L., Kwok, K.L., Dinstl, N., Deng, P.: TREC2003 Robust, HARD and QA Track Experiments using PIRCS (2003)
5. Dumais, S.T.: Latent Semantic Indexing (LSI) and TREC-2. In: Harman, D.K. (ed.) *Proceedings of TREC'2*, Gaithersburg. NIST, vol. 500-215, pp. 105–115 (1994)
6. Martínez-Santiago, F., Ureña, L.A., Martín, M.: A merging strategy proposal: two step retrieval status value method. *Information Retrieval* 9(1), 71–93 (2006)
7. Porter, M.F.: An algorithm for suffix stripping. *Program* 14, 130–137 (1980)
8. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a way of life: Okapi at TREC. *Information Processing and Management* 1, 95–108 (2000)
9. Savoy, J.: Cross-Language Information Retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management* 39, 75–115 (2003)
10. Llopis, F., García Puigcerver, H., Cano, M., Toral, A., Espi, H.: IR-n System, a Passage Retrieval Architecture. In: Sojka, P., Kopeček, I., Pala, K. (eds.) *TSD 2004. LNCS (LNAI)*, vol. 3206, pp. 57–64. Springer, Heidelberg (2004)
11. Callan, J.P., Lu, Z., Croft, W.B.: Searching distributed collections with inference networks. In: *Proceedings of the 18th International Conference of the ACM SIGIR'95*, pp. 21–28. The ACM Press, New York (1995)
12. Calve, A., Savoy, J.: Database merging strategy based on logistic regression. *Information Processing and Management* 36, 341–359 (2000)
13. Powell, A.L., French, J.C., Callan, J., Connell, M., Viles, C.L.: The impact of database selection on distributed searching. In: *Proceedings of the 23rd International Conference of the ACM-SIGIR'2000*, pp. 232–239. ACM Press, New York (2000)
14. Voorhees, E., Gupta, N.K., Johnson-Laird, B.: The collection fusion problem. In: Harman, D.K. (ed.) *Proceedings of the 3rd Text Retrieval Conference TREC-3*, National Institute of Standards and Technology, Special Publication, vol. 500-225, pp. 95–104 (1995)

Robust Ad-Hoc Retrieval Experiments with French and English at the University of Hildesheim

Thomas Mandl, René Hackl, and Christa Womser-Hacker

University of Hildesheim, Information Science
Marienburger Platz 22, D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. This paper reports on experiments submitted for the robust task at CLEF 2006 ad intended to provide a baseline for other runs for the robust task. We applied a system previously tested for ad-hoc retrieval. Runs for monolingual English and French were submitted. Results on both training as well as test topics are reported. Only for French, positive results above 0.2 MAP were achieved.

1 Introduction

We intended to provide a base line for the robust task at CLEF 2006. Our system applied to ad-hoc CLEF 2005 data [2] is an adaptive fusion system based on the MIMOR model [4]. For the base line experiments, we solely optimized blind relevance feedback (BRF) parameters based on a strategy developed by Carpineto et al. [1]. The basic retrieval engine is Lucene.

2 CLEF Retrieval Experiments with the MIMOR Approach

Two runs for the English and two for the French monolingual data were submitted. The results for both test and training topics are shown in table 1 and 2, respectively.

Table 1. Results for Submitted Monolingual Runs

<i>Run</i>	<i>Language</i>	<i>Stemming</i>	<i>BRF</i> <i>(docs. - terms)</i>	<i>GeoAve</i>	<i>MAP</i>
uhienmo1	English	Lucene	5-30	0.01%	7.98%
uhienmo2	English	Lucene	15-30	0.01%	7.12%
uhifmo1	French	Lucene	5-30	5.76%	28.50%
uhifmo2	French	Lucene	15-30	6.25%	29.85%

Only the runs for French have reached a competitive level of above 0.2 MAP. The results for the geometric average for the English topics are worse, because low performance for several topics led to a sharp drop in the performance according to this measure.

Table 2. Result for Training Topics for Submitted Monolingual Runs

<i>Run</i>	<i>Language</i>	<i>Stemming</i>	<i>BRF</i> (<i>docs. - terms</i>)	<i>GeoAve</i>	<i>MAP</i>
uhienmo1	English	Lucene	5-30	0.01%	7.16%
uhienmo2	English	Lucene	15-30	0.01%	6.33%
uhifmo1	French	Lucene	5-30	8.58%	25.26%
uhifmo2	French	Lucene	15-30	9.88%	28.47%

3 Future Work

For future experiments, we intend to exploit the knowledge on the impact of named entities on the retrieval process [5] as well as selective relevance feedback strategies in order to improve robustness [3].

References

- 1 Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An Information-Theoretic Approach to Automatic Query Expansion. *ACM Transactions on Information Systems* 19(1), 1–27 (2001)
- 2 Hackl, R., Mandl, T., Womser-Hacker, C.: Mono- and Cross-lingual Retrieval Experiments at the University of Hildesheim. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004*. LNCS, vol. 3491, pp. 165–169. Springer, Heidelberg (2005)
- 3 Kwok, K.L.: An Attempt to Identify Weakest and Strongest Queries. In: *ACM SIGIR 2005 Workshop: Predicting Query Difficulty - Methods and Applications*. Salvador - Bahia - Brazil (August 19, 2005), <http://www.haifa.ibm.com/sigir05-qp/papers/kwok.pdf>
- 4 Mandl, T., Womser-Hacker, C.: A Framework for long-term Learning of Topical User Preferences in Information Retrieval. *New Library World* 105(5/6), 184–195 (2004)
- 5 Mandl, T., Womser-Hacker, C.: The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: *Applied Computing 2005: Proc. ACM SAC Symposium on Applied Computing (SAC)*. Information Access and Retrieval (IAR) Track, Santa Fe, New Mexico, USA, March 13–17, 2005, pp. 1059–1064 (2005)

Comparing the Robustness of Expansion Techniques and Retrieval Measures

Stephen Tomlinson

Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
<http://www.opentext.com/>

Abstract. We investigate which retrieval measures successfully discern robustness gains in the monolingual (Bulgarian, French, Hungarian, Portuguese and English) information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2006. In all 5 of our experiments with blind feedback (a technique known to impair robustness across topics), the mean scores of the Average Precision, Geometric MAP and Precision@10 measures increased (and most of these increases were statistically significant), implying that these measures are not suitable as robust retrieval measures. In contrast, we found that measures based on just the first relevant item, such as a Generalized Success@10 measure, successfully discerned some robustness gains, particularly the robustness advantage of expanding Title queries by using the Description field instead of blind feedback.

1 Introduction

The goal of robust retrieval [9] is to reduce the likelihood of very poor results. The popular mean average precision (MAP) measure of ad hoc retrieval is considered unsuitable as a robustness measure; for example, blind feedback is a technique that typically boosts MAP but is considered bad for robustness because of its tendency to “not help (and frequently hurt) the worst performing topics” [9]. Success@10 (S10) has been suggested as a “direct measure” of robustness, but it “has the drawback of being a very coarse measure” [9].

Recently, we have investigated a Generalized Success@10 (GS10) measure as a robustness measure, and we have found that blind feedback can produce declines in it even when there is an increase in MAP, both in our own experiments [8][5][7], and in 7 other groups’ blind feedback experiments at the 2003 RIA Workshop [6]; many of these results were statistically significant in both directions. In [6], we investigated 30 retrieval measures across the 7 RIA systems, and the pattern was that measures with most of their weight on the rank of the first relevant item (such as reciprocal rank (MRR) and Success@10) tended to decline with blind feedback, while measures with most of their weight on additional relevant items (such as Precision@10 (P10) and MAP) tended to increase with blind feedback. The Generalized Success@10 measure was the most reliable of the 30 measures

at producing a statistically significant decline from blind feedback, suggesting that it was the best choice of those 30 measures as a robustness measure.

Recently, an experimental “Geometric MAP” (GMAP) measure was suggested as a robustness measure [10]. However, in our experiments in [7] and also in the RIA experiments reported in [6], we found that blind feedback produced statistically significant increases in GMAP and no statistically significant declines. Hence we do not believe that Geometric MAP is suitable as a robustness measure.

In this paper, we investigate whether the recent findings for blind feedback hold on the new data from the CLEF 2006 Ad-Hoc Track collections. Also, we look at a technique that should improve robustness, namely adding the Description field to a Title query, and check whether the measures give the expected results.

2 Retrieval Measures

“Primary recall” is retrieval of the first relevant item for a topic. Primary recall measures reported in this paper include the following:

- *Generalized Success@30* (GS30): For a topic, GS30 is 1.024^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Generalized Success@10* (GS10): For a topic, GS10 is 1.08^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Success@n* (S@n): For a topic, Success@ n is 1 if a desired page is found in the first n rows, 0 otherwise. This paper lists Success@10 (S10) and Success@1 (S1) for all runs.
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found. “Mean Reciprocal Rank” (MRR) is the mean of the reciprocal ranks over all the topics.

Note: GS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$. (Similarly, GS30 is considered a generalization of Success@30 because it rounds to 1 for $r \leq 30$ and to 0 for $r > 30$.) GS10 was introduced as “First Relevant Score” (FRS) in [8].

“Secondary recall” is retrieval of the additional relevant items for a topic (after the first one). Secondary recall measures place most of their weight on these additional relevant items and include the following:

- *Precision@n*: For a topic, “precision” is the percentage of retrieved documents which are relevant. “Precision@ n ” is the precision after n documents have been retrieved. This paper lists Precision@10 (P10) for all runs.
- *Average Precision* (AP): For a topic, AP is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant

documents which are not retrieved). By convention, AP is based on the first 1000 retrieved documents for the topic. The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). “Mean Average Precision” (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).

- *Geometric MAP* (GMAP): GMAP (introduced in [10]) is based on “Log Average Precision” which for a topic is the natural log of the max of 0.00001 and the average precision. GMAP is the exponential of the mean log average precision.
- *GMAP’*: We also define a linearized log average precision measure (denoted GMAP’) which linearly maps the ‘log average precision’ values to the [0,1] interval. For statistical significance purposes, GMAP’ gives the same results as GMAP, and it has advantages such as that the individual topic differences are in the familiar -1.0 to 1.0 range and are on the same scale as the mean.

3 Experiments

The CLEF 2006 Ad-Hoc Track [2] had 5 monolingual tasks, herein denoted as “BG” (Bulgarian), “FR” (French), “HU” (Hungarian), “PT” (Portuguese) and “EN” (English).

For each of these tasks, we performed 3 runs, herein denoted as “t” (just the Title field of the topic was used), “te” (same as “t” except that blind feedback

Table 1. Mean Scores of Monolingual Ad Hoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	GMAP	MAP
BG-td	0.903	0.829	44/50	0.648	26/50	0.308	0.172	0.285
BG-te	0.858	0.742	38/50	0.537	21/50	0.314	0.148	0.291
BG-t	0.832	0.724	38/50	0.513	19/50	0.282	0.108	0.261
FR-td	0.968	0.921	47/49	0.781	33/49	0.445	0.294	0.387
FR-te	0.940	0.861	45/49	0.674	28/49	0.433	0.266	0.390
FR-t	0.936	0.857	44/49	0.702	30/49	0.390	0.233	0.352
HU-td	0.954	0.881	45/48	0.675	26/48	0.377	0.203	0.298
HU-te	0.923	0.820	42/48	0.590	21/48	0.392	0.188	0.309
HU-t	0.911	0.817	43/48	0.582	20/48	0.354	0.128	0.267
PT-td	0.961	0.910	49/50	0.739	30/50	0.574	0.314	0.426
PT-te	0.927	0.874	46/50	0.700	29/50	0.568	0.248	0.420
PT-t	0.928	0.872	46/50	0.695	28/50	0.542	0.220	0.391
EN-td	0.957	0.885	46/49	0.683	27/49	0.404	0.306	0.409
EN-te	0.912	0.815	41/49	0.633	25/49	0.396	0.258	0.404
EN-t	0.910	0.814	41/49	0.643	26/49	0.378	0.229	0.371

Table 2. Impact of Adding the Description to Title-only Queries

Expt	Δ GS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-td-t	0.071	(0.010, 0.133)	22-7-21	0.78 (308), 0.76 (357), -0.64 (322)
FR-td-t	0.032	(-0.001, 0.064)	12-9-28	0.58 (309), 0.43 (325), -0.06 (347)
HU-td-t	0.044	(-0.004, 0.091)	15-9-24	0.85 (309), 0.52 (358), -0.25 (369)
PT-td-t	0.032	(0.000, 0.064)	17-9-24	0.49 (327), 0.49 (346), -0.08 (344)
EN-td-t	0.046	(0.012, 0.081)	18-9-22	0.53 (338), 0.43 (343), -0.09 (346)
Δ GS10				
BG-td-t	0.105	(0.024, 0.186)	22-7-21	0.99 (308), 0.99 (357), -0.62 (322)
FR-td-t	0.064	(0.003, 0.125)	12-9-28	0.94 (309), 0.84 (325), -0.16 (347)
HU-td-t	0.065	(-0.005, 0.135)	15-9-24	0.91 (358), 0.63 (321), -0.42 (369)
PT-td-t	0.038	(-0.003, 0.079)	17-9-24	0.59 (346), 0.48 (327), -0.21 (344)
EN-td-t	0.071	(0.015, 0.127)	18-9-22	0.91 (338), 0.54 (349), -0.23 (346)
Δ MRR				
BG-td-t	0.135	(0.038, 0.232)	22-7-21	0.98 (308), 0.98 (357), -0.67 (303)
FR-td-t	0.078	(-0.023, 0.180)	12-9-28	0.97 (309), 0.96 (325), -0.50 (318)
HU-td-t	0.092	(-0.004, 0.189)	15-9-24	0.97 (358), 0.89 (365), -0.75 (320)
PT-td-t	0.044	(-0.052, 0.140)	17-9-24	0.75 (318), 0.75 (350), -0.75 (303)
EN-td-t	0.040	(-0.056, 0.136)	18-9-22	0.97 (338), 0.91 (349), -0.67 (307)
Δ P10				
BG-td-t	0.026	(-0.026, 0.078)	18-13-19	0.50 (361), 0.50 (364), -0.50 (314)
FR-td-t	0.055	(0.006, 0.104)	18-9-22	0.40 (328), 0.40 (303), -0.40 (316)
HU-td-t	0.023	(-0.027, 0.073)	19-8-21	-0.60 (315), 0.40 (354), 0.40 (372)
PT-td-t	0.032	(-0.022, 0.086)	19-13-18	0.80 (337), 0.40 (323), -0.40 (304)
EN-td-t	0.027	(-0.018, 0.071)	15-13-21	0.50 (338), 0.40 (349), -0.20 (305)
Δ GMAP'				
BG-td-t	0.041	(0.009, 0.072)	34-15-1	0.44 (324), 0.43 (308), -0.13 (315)
FR-td-t	0.020	(0.003, 0.037)	30-19-0	0.21 (317), 0.20 (309), -0.11 (315)
HU-td-t	0.040	(0.006, 0.074)	36-9-3	0.72 (309), 0.32 (358), -0.16 (315)
PT-td-t	0.031	(0.001, 0.061)	30-19-1	0.59 (320), 0.34 (323), -0.12 (315)
EN-td-t	0.025	(0.010, 0.040)	25-21-3	0.22 (338), 0.13 (313), -0.04 (324)
Δ MAP				
BG-td-t	0.024	(-0.015, 0.063)	34-15-1	-0.35 (373), -0.31 (315), 0.32 (353)
FR-td-t	0.034	(0.001, 0.068)	30-19-0	-0.28 (315), 0.25 (341), 0.27 (311)
HU-td-t	0.031	(0.009, 0.054)	36-9-3	0.22 (318), -0.19 (353), -0.20 (315)
PT-td-t	0.035	(0.000, 0.070)	30-19-1	0.46 (341), 0.35 (337), -0.32 (315)
EN-td-t	0.038	(-0.002, 0.079)	25-21-3	0.66 (341), 0.44 (338), -0.25 (324)

(based on the first 3 rows of the “t” query) was used to expand the query), and “td” (same as “t” except that the Description field was additionally used).

The details of indexing, ranking and blind feedback were essentially the same as last year [8]. For Bulgarian and Hungarian, we thank Jacques Savoy for providing stemmers and updated stop word lists at [4].

Table 3. Impact of Blind Feedback Expansion on Title-only Queries

Expt	Δ GS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-te-t	0.026	(−0.012, 0.065)	14-11-25	0.83 (324), 0.34 (374), −0.17 (357)
FR-te-t	0.004	(−0.010, 0.017)	7-8-34	0.19 (325), 0.14 (320), −0.09 (336)
HU-te-t	0.012	(−0.022, 0.046)	8-9-31	0.77 (309), −0.11 (374), −0.12 (357)
PT-te-t	−0.001	(−0.012, 0.009)	6-8-36	0.17 (323), −0.05 (343), −0.16 (327)
EN-te-t	0.002	(−0.015, 0.018)	11-7-31	−0.21 (322), 0.14 (343), 0.19 (309)
Δ GS10				
BG-te-t	0.018	(−0.011, 0.047)	14-11-25	0.54 (324), 0.28 (374), −0.16 (352)
FR-te-t	0.005	(−0.020, 0.029)	7-8-34	0.24 (325), 0.22 (317), −0.21 (347)
HU-te-t	0.003	(−0.022, 0.028)	8-9-31	0.43 (309), −0.16 (375), −0.23 (357)
PT-te-t	0.002	(−0.015, 0.018)	6-8-36	0.25 (323), −0.11 (325), −0.14 (343)
EN-te-t	0.001	(−0.022, 0.023)	11-7-31	0.26 (309), −0.18 (322), −0.25 (318)
Δ MRR				
BG-te-t	0.024	(−0.018, 0.067)	14-11-25	0.50 (360), 0.50 (363), −0.50 (365)
FR-te-t	−0.029	(−0.082, 0.025)	7-8-34	−0.67 (340), −0.67 (312), 0.50 (328)
HU-te-t	0.007	(−0.041, 0.056)	8-9-31	−0.50 (314), −0.50 (364), 0.50 (372)
PT-te-t	0.006	(−0.051, 0.063)	6-8-36	−0.67 (343), −0.50 (326), 0.50 (312)
EN-te-t	−0.010	(−0.061, 0.040)	11-7-31	−0.67 (345), 0.50 (329), 0.50 (325)
Δ P10				
BG-te-t	0.032	(0.006, 0.058)	13-4-33	0.40 (364), 0.30 (319), −0.10 (304)
FR-te-t	0.043	(0.006, 0.080)	19-7-23	0.40 (345), 0.40 (328), −0.20 (346)
HU-te-t	0.038	(0.008, 0.067)	13-3-32	0.40 (319), 0.30 (354), −0.20 (305)
PT-te-t	0.026	(−0.006, 0.058)	17-5-28	−0.40 (316), −0.20 (331), 0.30 (342)
EN-te-t	0.018	(−0.012, 0.049)	15-4-30	0.30 (328), −0.30 (305), −0.30 (344)
Δ GMAP'				
BG-te-t	0.028	(−0.004, 0.059)	35-14-1	0.78 (324), 0.08 (310), −0.04 (368)
FR-te-t	0.012	(0.006, 0.017)	42-7-0	0.06 (323), 0.04 (307), −0.05 (312)
HU-te-t	0.034	(0.000, 0.067)	40-5-3	0.78 (309), 0.10 (368), −0.02 (357)
PT-te-t	0.010	(0.003, 0.017)	38-10-2	0.10 (323), 0.08 (337), −0.04 (327)
EN-te-t	0.010	(0.003, 0.018)	37-9-3	0.12 (309), 0.05 (341), −0.04 (316)
Δ MAP				
BG-te-t	0.029	(0.018, 0.041)	35-14-1	0.15 (364), 0.14 (319), −0.02 (373)
FR-te-t	0.038	(0.021, 0.055)	42-7-0	0.25 (341), 0.12 (328), −0.14 (312)
HU-te-t	0.042	(0.025, 0.059)	40-5-3	0.28 (319), 0.18 (364), −0.01 (321)
PT-te-t	0.029	(0.014, 0.044)	38-10-2	0.15 (341), 0.13 (306), −0.14 (332)
EN-te-t	0.033	(0.018, 0.049)	37-9-3	0.20 (336), 0.19 (341), −0.10 (344)

Table 1 lists the mean scores of the 15 runs.

Table 2 shows that expanding the Title queries by adding the Description field increased the mean score for all investigated measures (GS30, GS10, MRR, P10, GMAP and MAP), including at least one statistically significant increase for each measure. Adding the Description is a robust technique that can sometimes improve a poor result from just using the Title field.

Table 4. Comparison of Expansion Techniques for Title-only Queries

Expt	Δ GS30	95% Conf	vs.	3 Extreme Diffs (Topic)
BG-td-te	0.045	(−0.021, 0.112)	18-10-22	0.93 (357), 0.80 (308), −0.65 (324)
FR-td-te	0.028	(−0.002, 0.059)	15-8-26	0.64 (309), 0.25 (325), −0.10 (320)
HU-td-te	0.032	(−0.002, 0.065)	15-8-25	0.42 (358), 0.35 (352), −0.27 (369)
PT-td-te	0.034	(−0.002, 0.069)	18-10-22	0.65 (327), 0.51 (346), −0.07 (303)
EN-td-te	0.045	(0.010, 0.080)	16-9-24	0.49 (338), 0.38 (322), −0.11 (346)
Δ GS10				
BG-td-te	0.087	(0.004, 0.170)	18-10-22	1.00 (357), 0.99 (308), −0.54 (324)
FR-td-te	0.060	(0.005, 0.114)	15-8-26	0.96 (309), 0.60 (325), −0.17 (320)
HU-td-te	0.062	(−0.005, 0.128)	15-8-25	0.83 (358), 0.69 (321), −0.47 (369)
PT-td-te	0.037	(−0.003, 0.076)	18-10-22	0.59 (346), 0.50 (327), −0.21 (303)
EN-td-te	0.070	(0.012, 0.128)	16-9-24	0.88 (338), 0.51 (313), −0.30 (346)
Δ MRR				
BG-td-te	0.111	(0.007, 0.214)	18-10-22	0.99 (357), 0.99 (308), −0.67 (303)
FR-td-te	0.107	(0.001, 0.212)	15-8-26	0.98 (309), 0.92 (325), −0.50 (337)
HU-td-te	0.085	(−0.022, 0.192)	15-8-25	0.96 (358), 0.90 (365), −0.75 (372)
PT-td-te	0.038	(−0.061, 0.137)	18-10-22	0.75 (350), 0.75 (318), −0.75 (303)
EN-td-te	0.051	(−0.037, 0.139)	16-9-24	0.97 (338), 0.90 (349), −0.67 (307)
Δ P10				
BG-td-te	−0.006	(−0.058, 0.046)	15-17-18	−0.50 (314), −0.40 (373), 0.50 (361)
FR-td-te	0.012	(−0.038, 0.063)	15-15-19	−0.50 (345), 0.40 (303), 0.40 (340)
HU-td-te	−0.015	(−0.068, 0.039)	16-14-18	−0.60 (315), −0.60 (311), 0.40 (372)
PT-td-te	0.006	(−0.049, 0.061)	13-17-20	0.60 (316), 0.60 (337), −0.30 (336)
EN-td-te	0.008	(−0.037, 0.053)	14-19-16	0.50 (338), 0.30 (344), −0.20 (328)
Δ GMAP'				
BG-td-te	0.013	(−0.016, 0.042)	25-25-0	0.37 (308), 0.26 (357), −0.34 (324)
FR-td-te	0.009	(−0.006, 0.024)	22-25-2	0.18 (317), 0.18 (309), −0.11 (315)
HU-td-te	0.007	(−0.013, 0.026)	23-22-3	0.30 (358), −0.13 (368), −0.18 (315)
PT-td-te	0.021	(−0.008, 0.049)	23-25-2	0.59 (320), 0.24 (323), −0.13 (315)
EN-td-te	0.015	(−0.001, 0.030)	26-20-3	0.24 (338), 0.13 (313), −0.07 (336)
Δ MAP				
BG-td-te	−0.006	(−0.041, 0.030)	25-25-0	−0.33 (373), −0.31 (315), 0.30 (353)
FR-td-te	−0.004	(−0.037, 0.030)	22-25-2	−0.28 (315), −0.27 (345), 0.23 (317)
HU-td-te	−0.011	(−0.040, 0.018)	23-22-3	−0.29 (311), −0.27 (315), 0.16 (324)
PT-td-te	0.006	(−0.029, 0.041)	23-25-2	−0.35 (315), −0.34 (313), 0.31 (341)
EN-td-te	0.005	(−0.034, 0.045)	26-20-3	0.47 (341), 0.45 (338), −0.30 (324)

Table 3 shows that expanding the Title queries via blind feedback of the first 3 rows did not produce any statistically significant increases for the primary recall measures (GS30, GS10, MRR), even though it produced statistically significant increases for the secondary recall measures (P10, GMAP, MAP). Blind feedback is not a robust technique in that it is unlikely to improve poor results. (In a larger experiment, we would expect the primary recall measures to show statistically significant decreases, like we saw for Bulgarian last year [8].)

Table 4 compares the results of the two title-expansion approaches. For each primary recall measure (GS30, GS10, MRR), there is at least one positive statistically significant difference, reflecting the robustness of expanding using the Description instead of blind feedback. However, there are no statistically significant differences in the secondary recall measures (P10, GMAP, MAP); these measures do not discern the higher robustness of the “td” run compared to the “te” run.

4 Conclusions and Future Work

For all 5 blind feedback experiments reported in this paper, the mean scores for MAP, GMAP and P10 were up with blind feedback, and most of these increases were statistically significant. As blind feedback is known to be bad for robustness (because of its tendency to “not help (and frequently hurt) the worst performing topics” [9]), we continue to conclude that none of these 3 measures should be used as robustness measures.

Note that [3] recently made the (unsupported) claim that for GMAP, “blind feedback is often found to be detrimental”. Our results for GMAP, both in this paper and in past experiments ([7][6]) do not support such a claim.

Measures based on just the first relevant item (i.e. primary recall measures such as GS30 and GS10) reflect robustness. In this paper, we found in particular that these measures discerned the robustness advantage of expanding Title queries by using the Description field instead of blind feedback, while the secondary recall measures (MAP, GMAP, P10) did not.

[1] gives a theoretical explanation for why different retrieval approaches are superior when seeking just one relevant item instead of several. In particular, it finds that when seeking just one relevant item, it can theoretically be advantageous to use *negative* pseudo-relevance feedback to encourage more diversity in the results.

Intuitively, a recall-oriented measure would be robust if it just counted distinct aspects of a topic, such as the “instance recall” metric described in [1]. (We presently have not done experiments with this measure.) However, such a metric requires special assessor effort to mark the distinct aspects (which was done for 20 topics for the TREC-6, 7 and 8 Interactive Tracks, according to [1]).

For binary assessments (relevant vs. not-relevant), incorporating extra relevant items into a metric apparently causes it to lose its robustness. Intuitively, the problem is that the metric starts to reward the retrieval of duplicated information. For example, Precision@10 can get a perfect 1.0 score by retrieving 10 similar relevant items; if a system returned 10 different interpretations of the query, it would be unlikely to get a perfect Precision@10 score, but it would be more likely to get credit for Success@10, particularly if the user wanted one of the less likely interpretations of the query.

To encourage more research in robust retrieval, probably the simplest thing the organizers of ad hoc tracks could do would be to use a measure based on just the first relevant item (e.g. GS10 or GS30) as the primary measure for the

ad hoc task. Participants would then find it detrimental to use the non-robust blind feedback technique, but potentially would be rewarded for finding ways of producing more diverse results.

References

1. Chen, H., Karger, D.R.: Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In: SIGIR 2006, pp. 429–436 (2006)
2. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. [In this volume]
3. Robertson, S.: On GMAP – and other transformations. In: CIKM 2006, pp. 78–83 (2006)
4. Savoy, J.: CLEF and Multilingual information retrieval resource page <http://www.unine.ch/info/clef/>
5. Tomlinson, S.: CJK Experiments with Hummingbird SearchServerTM at NTCIR-5. In: Proceedings of NTCIR-5 (2005)
6. Tomlinson, S.: Early Precision Measures: Implications from the Downside of Blind Feedback. In: SIGIR 2006, pp. 705–706 (2006)
7. Tomlinson, S.: Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServerTM at TREC 2005. In: Proceedings of TREC 2006 (2006)
8. Tomlinson, S.: European Ad Hoc Retrieval Experiments with Hummingbird SearchServerTM at CLEF 2005. In: Working Notes for the CLEF 2005 Workshop (2005)
9. Voorhees, E.M.: Overview of the TREC 2003 Robust Retrieval Track. In: Proceedings of TREC 2004 (2004)
10. Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track. In: Proceedings of TREC 2005 (2005)

Experiments with Monolingual, Bilingual, and Robust Retrieval

Jacques Savoy and Samir Abdou

Computer Science Department,
University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchatel, Switzerland
Jacques.Savoy@unine.ch, Samir.Abdou@unine.ch

Abstract. For our participation in the CLEF 2006 campaign, our first objective was to propose and evaluate a decompounding algorithm and a more aggressive stemmer for the Hungarian language. Our second objective was to obtain a better picture of the relative merit of various search engines for the French, Portuguese/Brazilian and Bulgarian languages. To achieve this we evaluated the test-collections using the Okapi approach, some of the models derived from the *Divergence from Randomness* (DFR) family and a language model (LM), as well as two vector-processing approaches. In the bilingual track, we evaluated the effectiveness of various machine translation systems for a query submitted in English and automatically translated into the French and Portuguese languages. After blind query expansion, the MAP achieved by the best single MT system was around 95% for the corresponding monolingual search when French was the target language, or 83% with Portuguese. Finally, in the robust retrieval task we investigated various techniques in order to improve the retrieval performance of difficult topics.

1 Introduction

During the last few years, the IR group at University of Neuchatel has been involved in designing, implementing and evaluating IR systems for various natural languages, including both European [1] and popular Asian [2] languages. In this context, our first objective was to promote effective monolingual IR for these languages. Our second aim was to design and evaluate effective bilingual search techniques (using a query-based translation approach), and our third objective was to propose effective multilingual IR systems.

The rest of this paper is organized as follows: Section 2 explains the principal features of different indexing and search strategies, and then evaluates them using the available corpora. The data fusion approaches used in our experiments and our official results are outlined in Section 3. Our bilingual experiments are presented and evaluated in Section 4. Finally, Section 5 presents our first experiments in the robust track, limited however to the French language.

2 Indexing and Searching Strategies

In order to obtain a broader view of the relative merit of various retrieval models, we first adopted the classical *tf · idf* weighting scheme (with cosine normalization). We then computed the inner product to measure similarity between documents and requests. Although various other indexing weighting schemes have been suggested, in our study we will only consider the IR model "Lnu" [3].

In addition to these two IR models based on the vector-space paradigm, we also considered probabilistic approaches such as the Okapi model [4]. As a second probabilistic approach, we implemented several models derived from the *Divergence from Randomness* (DFR) family [5]. The GL2 approach for example is based on the following equations:

$$w_{ij} = Inf_{ij}^1 \cdot Inf_{ij}^2 = Inf_{ij}^1 \cdot (1 - Prob_{ij}^2) \quad \text{with} \quad (1)$$

$$Prob_{ij}^2 = tfn_{ij} / (tfn_{ij} + 1) \quad (2)$$

$$tfn_{ij} = tf_{ij} \cdot \log_2 [1 + ((c \cdot mean\ dl) / l_i)] \quad (3)$$

$$Inf_{ij}^1 = -\log_2 [1/(1 + \lambda_j)] - tfn_{ij} \cdot \log_2 [\lambda_j/(1 + \lambda_j)], \quad \lambda_j = tc_j/n \quad (4)$$

where tc_j represents the number of occurrences of term t_j in the collection, l_i the length of document D_i , *mean dl* is the document mean length, and n the number of documents in the corpus. On the other hand, the PL2 model uses Equation 5 instead of Equation 4.

$$Inf_{ij}^1 = -\log_2 \left[\frac{e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}}}{tf_{ij}!} \right] \quad (5)$$

Finally, we considered an approach based on the language model (LM) [6] known as a non-parametric probabilistic model (the Okapi and DFR are viewed as parametric models). Probability estimates would thus be estimated directly, based on occurrence frequencies in document D_i or corpus C . Within this language model paradigm, various implementation and smoothing methods might be considered, and in this study we adopted a model which combines an estimate based on document ($P[t_j|D_i]$) and on corpus ($P[t_j|C]$) [6].

$$Prob[D_i|Q] = Prob[D_i] \prod_{t_j \in Q} [\lambda_j \cdot Prob[t_j|D_i] + (1 - \lambda_j) \cdot Prob[t_j|C]] \quad (6)$$

$$Prob[t_j|D_i] = tf_{ij}/l_i \quad \text{and} \quad Prob[t_j|C] = df_j/lc \quad \text{with} \quad lc = \sum_k df_k \quad (7)$$

where λ_j is a smoothing factor (constant for all indexing terms t_j , and fixed at 0.35) and lc the size of the corpus C .

During this evaluation campaign, we applied the stopword lists and stemmers that were used in our CLEF 2005 participation [7]. In our Bulgarian stopwords list however we corrected certain errors (including the removal of words having a

clear meaning and introduced by mistake in the suggested stopwords list). For the Hungarian collection, we automatically compounded long words (more than 6 characters) using our own algorithm [8]. In this experiment, compound words were replaced by their components in both documents and queries. This year, we tried to be more aggressive, adding 17 rules to our Hungarian stemmer in order to also remove certain derivational suffixes (e.g., "jelent" (to mean) and "jelentés" (meaning), or "tánc" (to dance) and "táncol" (dance)).

To measure the retrieval performance, we adopted the mean average precision (MAP) computed by the `trec_eval` system. Then we applied the bootstrap methodology [9] in order to statistically determine whether or not a given search strategy would be better than another. Thus, in the tables included in this paper we underline statistically significant differences resulting from the use of a two-sided non-parametric bootstrap test (significance level fixed at 5%).

We indexed the various collections using words as indexing units. Table 1 shows evaluations on our four probabilistic models, as well as two vector-space schemes. The best performances under given conditions are shown in bold type in this table, and they are used as a baseline for our statistical testing. The underlined results therefore indicate which MAP differences can be viewed as statistically significant when compared to the best system value. As shown in the top part of Table 1, the Okapi model was the best IR model for the French and Portuguese/Brazilian collections. For these two corpora however, MAP differences between the various probabilistic IR models were not always statistically significant. The DFR-GL2 model provided the best results for the Bulgarian collection, while for the Hungarian corpus, the DFR-PL2 approach resulted in the best performance.

Table 1. MAP of Single Searching Strategies

Query Model	Mean average precision			
	French TD	Portuguese TD	Bulgarian TD	Hungarian TD
	49 queries	50 queries	50 queries	48 queries
Okapi	0.4151	0.4118	0.2614	0.3392
DFR-PL2	0.4101	0.4147	N/A	0.3399
DFR-GL2	0.3988	0.4033	0.2734	0.3396
LM	0.3913	0.3909	0.2720	0.3344
Lnu-ltc	<u>0.3738</u>	<u>0.4212</u>	0.2663	0.3303
<i>tf idf</i>	<u>0.2606</u>	<u>0.2959</u>	<u>0.1898</u>	<u>0.2623</u>

It was observed that pseudo-relevance feedback (PRF or blind-query expansion) seemed to be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio's approach [3] with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add m terms extracted from the k best ranked documents from the original query. To evaluate this proposition, we used three IR

Table 2. MAP Before and After Blind-Query Expansion

Query TD Model	Mean average precision			
	French 49 queries	Portuguese 50 queries	Bulgarian 50 queries	Hungarian 48 queries
Okapi	0.4151	0.4118	0.2614	0.3392
k docs/ m terms	10/20 0.4222	10/20 0.4236	3/50 0.2833	3/10 0.3545
	10/30 0.4269	10/30 0.4361	5/50 0.2798	5/10 0.3513
	10/40 <u>0.4296</u>	10/40 0.4362	3/90 0.2854	5/15 0.3490
	10/50 0.4261	10/50 <u>0.4427</u>	5/90 0.2809	10/15 0.3492
DFR-GL2	0.3988	0.4033	0.2734	0.3396
k / m	10/50 <u>0.4356</u>	10/20 0.4141	5/10 <u>0.3327</u>	5/50 <u>0.4059</u>
LM	0.3913	0.3909	0.2720	0.3344
k / m	10/50 <u>0.4509</u>	10/30 <u>0.4286</u>	10/20 <u>0.3305</u>	3/40 <u>0.3855</u>

models and enlarged the query by the 10 to 90 terms extracted from the 3 to 10 best-ranked articles (see Table 2).

For the French corpus, the percentage of improvement varied from +3.5% (Okapi model, 0.4151 vs. 0.4296) to +15.2% (LM model, 0.3913 vs. 0.4509). For the Portuguese/Brazilian corpus, the increased enhancement was +2.7% (DFR-GL2 model, 0.4033 vs. 0.4141) to +9.6% (LM model, 0.3909 vs. 0.4286). For the Bulgarian language, the use of a blind query expansion improved the MAP from +9.2% (Okapi model, 0.2614 vs. 0.2854) to +21.7% (DFR-GL2 model, 0.2734 vs. 0.3327). Finally, with the Hungarian language, a blind query expansion might enhance retrieval effectiveness from +4.5% (Okapi model, 0.3392 vs. 0.3545) to +19.5% (DFR-GL2 model, 0.3396 vs. 0.4059).

3 Data Fusion and Official Results

It is assumed that combining different search models should improve retrieval effectiveness, due to the fact that different document representations might retrieve different pertinent items, thus increasing overall recall [10]. In this current study we combined two or three probabilistic models, representing both the parametric (Okapi and DFR) and non-parametric (LM) probabilistic approaches. To achieve this we evaluated various fusion operators (see [7] for a list of their precise descriptions) such as the Norm RSV, where a normalization procedure is applied before adding document scores computed by different search models.

Table 3 shows the exact specifications of our best performing official monolingual runs. In these experiments, we combined the Okapi, and the LM probabilistic models using the Z-score data fusion operator [7] for the French and Portuguese/Brazilian corpora. We obtained the best results when using the LM model combined with the DFR-GL2 model for the Bulgarian corpus or when combining the Okapi, DFR-PL2 and LM models for the Hungarian language.

Table 3. Description and MAP of our Best Official Monolingual Runs

Language	Index	Query	Model	Query exp.	MAP	comb. MAP
French	word	TD	Okapi	10 docs/60 terms	0.4275	Norm RSV
UniNEfr3	word	TD	LM	10 docs/30 terms	0.4460	0.4559
Portuguese	word	TD	Okapi	10 docs/80 terms	0.4276	Z-score
UniNEpt1	word	TD	LM	10 docs/50 terms	0.4403	0.4552
Bulgarian	word	TD	LM	5 docs/40 terms	0.3201	Z-score
UniNEbg2	4-gram	TD	GL2	10 docs/90 terms	0.2941	0.3314
Hungarian	word	TD	PL2	3 docs/40 terms	0.3794	Z-score
	word	TD	LM	3 docs/70 terms	0.3815	
UniNEhu2	4-gram	TD	Okapi	3 docs/100 terms	0.3870	0.4308

4 Bilingual Information Retrieval

Due to a time constraint, we limited our participation in the bilingual track to the French and Portuguese/Brazilian languages. Moreover, as the query submission language we chose English, which was automatically translated into the two other languages, using ten freely available machine translation (MT) systems (listed in the first column of Table 4).

Table 4. MAP of Various Translation Devices (Okapi model)

	Mean average precision (% monolingual)	
TD queries	French	Portuguese
Model	49 queries	50 queries
Manual & PRF (10/40)	0.4296	0.4389
AlphaWorks	0.3378 (78.6%)	N/A
AppliedLanguage	0.3726 (86.7%)	0.3077 (70.1%)
BabelFish	0.3771 (87.8%)	0.3092 (70.4%)
FreeTranslation	0.3813 (88.8%)	0.3356 (76.5%)
Google	0.3754 (87.4%)	0.3070 (69.9%)
InterTrans	0.2761 (64.3%)	0.3343 (76.2%)
Online	0.3941 (91.8%)	0.3677 (83.3%)
Reverso	0.4081 (95.0%)	0.3531 (80.5%)
Systran	N/A	0.3077 (70.1%)
WorldLingo	0.3832 (89.2%)	0.3091 (70.4%)

The results of experiments are shown in Table 4, indicating that Reverso provided the best translation for the French collection and Online for the Portuguese corpus. With the FreeTranslation system, these three MT systems usually obtained satisfactory retrieval performances for both languages. For French, the WorldLingo, BabelFish, or Google translation systems also worked well.

Finally, Table 5 lists the parameter settings used for our best performing official runs in the bilingual task. Based on our previous experiments [7], we

Table 5. Description and MAP of our Best Official Bilingual Runs

From EN to ...	French 49 queries	French 49 queries	Portuguese 50 queries	Portuguese 50 queries
IR 1 (doc/term)	PL2 (10/30)	PL2 (10/30)	I(n)L2 (10/40)	GL2 (10/40)
IR 2 (doc/term)	LM (10/30)	Okapi (10/60)	LM (10/30)	Okapi (10/80)
IR 3 (doc/term)		LM (10/50)		LM (10/40)
Data fusion	Z-score	Z-score	Round-robin	Z-score
Translation tools	BabelFish & Reverso	Reverso & Online	Prompt & Free & Online	Prompt & Free
MAP	0.4278	0.4256	0.4114	0.4138
Run name	UniNEbifr1	UniNEbifr2	UniNEbipt1	UniNEbipt2

first concatenated two or three query translations obtained by different freely available translation tools. Before combining the result lists obtained by various search models, we automatically expanded the translated queries using a pseudo-relevance feedback method (Rocchio), as described in Table 5.

5 French Robust Track

The aim of this track is to analyze and to improve IR system performance when processing difficult topics [11], or queries from previous CLEF evaluation campaigns that have poor MAP. The goal of the robust track is therefore to explore various methods of building a search system that will perform "reasonably well" for all queries. In real systems this is an important concern, particularly when evaluating situations where the search engine returns unexpected results or "silly" responses to users.

In this track we reused queries created and evaluated during the CLEF 2001, 2002, and 2003 campaigns, with topic collections being the same for the most part. Moreover, the organizers arbitrarily divided this query set into a training set (60 queries) and a test set (100 queries). In the latter, 9 queries did not in fact obtain any relevant items and thus the test set only contained the 91 remaining queries. When analyzing this sample, we found that the mean number of relevant items per query was 24.066 (median: 14, minimum: 1, maximum: 177, standard deviation: 30.78).

When using the MAP to measure the retrieval effectiveness, all observations (queries) had the same importance. This arithmetic mean thus did not really penalize incorrect answers. Thus, Voorhess [11] and other authors suggested replacing the arithmetic mean (MAP) with the geometric mean (GMAP), in order to assign more importance to the poor performances obtained by difficult topics (both measures are depicted in Table 6).

Given our past experience, we decided to search the French collection using the three probabilistic models described in Section 2, as well as blind query expansion. As depicted in Table 6, the MAP resulting from these three models when applying the TD or T query formulations are relatively similar. These

Table 6. Description of our Official Robust Runs (French corpus, 91 queries)

Run name	Query	Model	Query exp.	MAP	comb. MAP	GMAP
UniNEfrr1	TD	Okapi	5 docs / 15 terms	0.5035	Round-Robin	0.3889
	TD	GL2	3 docs / 30 terms	0.5014		
	TD	LM	10 docs / 15 terms	0.5095		
UniNEfrr2	T	Okapi	3 docs / 10 terms	0.4058	Z-score	0.2376
	T	GL2	5 docs / 30 terms	0.4029		
	T	LM	5 docs / 10 terms	0.4137		
UniNEfrr3	TD	GL2	3 docs / 30 terms & Yahoo!.fr	0.4607		0.2935

result lists were then combined using a data fusion operator. This procedure was applied to two of our official runs, namely UniNEfrr1 with TD query, and UniNEfrr2 with T query (complete description given in Table 6).

For the last run (UniNEfrr3) we submitted the topic titles to Yahoo.fr search engine. The response page contained ten references, plus a short description. We then extracted these ten short textual descriptions and added them to the original query. The expanded query was then sent to our search model in order to hopefully obtain better results. With the TD query formulation, the mean number of distinct search terms was 7.51, and when including the first ten references retrieved by Yahoo!, the average value increased to 115.46 (meaning that we have added, in mean, 108 new search terms). This massive query expansion did not prove to be effective (MAP: 0.5014 before, and 0.4607 after query expansion using Yahoo! snippets) and in our efforts to improve retrieval performance we would most certainly need to include a term selection procedure.

6 Conclusion

During the CLEF 2006 evaluation campaign, we proposed a more effective search strategy for the Hungarian language. In an attempt to remove the more frequent derivational suffixes were applied a more aggressive stemmer, and we also evaluated an automatic decompounding scheme. Combining different indexing and retrieval schemes seems to be really effective for this language, although more processing time and disk space are required.

For the French, Portuguese/Brazilian and Bulgarian languages, we used the same stopword lists and stemmers developed during the previous years. In order to enhance retrieval performance, we implemented an IR model based on the language model and suggested a data fusion approach based on the Z-score, after applying a blind query expansion.

In the bilingual task, the freely available translation tools performed at a reasonable level for both the French and Portuguese languages (based on the best translation tool, compared to the monolingual search the MAP is around 95% for French and 83% for Portuguese/Brazilian). Finally, in the robust retrieval task,

we investigated some of the difficult topics, plus various methods that might be implemented to improve retrieval effectiveness.

Acknowledgments. The authors would like to thank Pierre-Yves Berger for his help in translating the English topics and in using the Yahoo.fr search engine. This research was supported by the Swiss NSF under Grants #200020-103420 and #200021-113273.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. *IR Journal* 7, 121–148 (2004)
2. Savoy, J.: Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. *ACM Transactions on Asian Languages Information Processing* 4, 163–189 (2005)
3. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: *Proceedings TREC-4*, Gaithersburg, pp. 25–48 (1996)
4. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management* 36, 95–108 (2002)
5. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems* 20, 357–389 (2002)
6. Hiemstra, D.: Using Language Models for Information Retrieval. Ph.D. Thesis (2000)
7. Savoy, J., Berger, P.-Y.: Monolingual, Bilingual, and GIRT Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 131–140. Springer, Heidelberg (2006)
8. Savoy, J.: Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic Models for Effective Monolingual Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) *CLEF 2003*. LNCS, vol. 3237, pp. 322–336. Springer, Heidelberg (2004)
9. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management* 33, 495–512 (1997)
10. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. *IR Journal* 1, 151–173 (1999)
11. Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track. In: *Proceedings TREC-2004*, Gaithersburg, pp. 70–79 (2005)

Local Query Expansion Using Terms Windows for Robust Retrieval

Angel F. Zazo, Jose L. Alonso Berrocal, and Carlos G. Figuerola

REINA Research Group – University of Salamanca
C/ Francisco Vitoria 6-16, 37008 Salamanca, Spain
<http://reina.usal.es>

Abstract. This paper describes our work at CLEF 2006 Robust task. This is an ad-hoc task that explores methods for stable retrieval by focusing on poorly performing topics. We have participated in all subtasks: monolingual (English, French, Italian and Spanish), bilingual (Italian to Spanish) and multilingual (Spanish to [English, French, Italian and Spanish]). In monolingual retrieval we have focused our effort on local query expansion, i.e. using only the information from retrieved documents, not from the complete document collection or external corpora, such as the Web. Some local expansion techniques were applied for training topics. Regarding robustness the most effective one was the use of co-occurrence based thesauri, which were constructed using co-occurrence relations in windows of terms, not in complete documents. This is an effective technique that can be easily implemented by tuning only a few parameters. In bilingual and multilingual retrieval experiments several machine translation programs were used to translate topics. For each target language, translations were merged before performing a monolingual retrieval. We also applied the same local expansion technique. In multilingual retrieval, weighted max-min normalization was used to merge lists. In all the subtasks in which we participated our mandatory runs (using title and description fields of the topics) obtained very good rankings. Runs with short queries (only title field) also obtained high MAP and GMAP values using the same expansion technique.

1 Introduction

This year our research group has participated in two tracks at CLEF 2006: Ad-hoc Robust Task and Web Track. This paper is limited to the former. For the latter, please see the relevant paper published in this volume.

Robust retrieval tries to obtain stable performance over all topics by focusing on poorly performing topics. Robust tracks were carried out at TREC 2003, 2004 and 2005 for monolingual retrieval [1,2,3], but not for cross-language information retrieval (CLIR). The goal of CLEF 2006 Robust Task is to consider not only monolingual, but also bilingual and multilingual retrieval. This is essentially an ad-hoc task, and it uses test collections developed at CLEF 2001 through CLEF 2003. The collections contain data in six languages, Dutch, English, German, French, Italian and Spanish, and a set of 160 topics, divided into *test* and

training topics. For a complete description of this task, please, see the CLEF 2006 Ad-hoc Track Overview, also published in this volume. Our research group has participated in all the subtasks of CLEF 2006 Robust Task: monolingual (English, French, Italian and Spanish), bilingual (Italian to Spanish) and multilingual (Spanish to [English, French, Italian and Spanish]). For each subtask and topic language two runs were submitted, one with the title and description fields of the topics (mandatory) and the other only with the title field.

The users of an information retrieval system ignore concepts such as average precision, recall, etc. They simply use it, and they tend to remember failures rather than successes, since failures are the decisive factor to use a system again. The system's robustness ensures that all topics obtain minimum effectiveness levels. In information retrieval the mean of the average precision (MAP) is used to measure systems' performance. But, poorly performing topics have little influence on MAP. At TREC, geometric average (GMAP), rather than MAP, turned out to be the most stable evaluation method for robustness [2]. The GMAP has the desired effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between higher scores. Nevertheless, at the CLEF 2006 Workshop the submitted runs showed high correlations between MAP and GMAP. Perhaps other used-related measurements, such as P10 or GS10, should be applied (see also the paper by Stephen Tomlinson in this volume).

At TREC, query expansion using both the document collection and external corpora (Web and other document collections) was the best approach for robust retrieval. But in all cases, many parameters had to be adjusted. We have focused our work on local query expansion, i.e. using only the information from retrieved documents, not from the complete document collection or external corpora, such as the Web. We applied two query expansion techniques: blind relevance feedback (BRF) and co-occurrence based thesauri (in this case using as co-occurrence units both the complete document and windows of terms).

In the experiments, it was only permitted to use relevance assessments of training topics. The topics of the robust task came from CLEF 2001 through CLEF 2003, but the document collections only came from CLEF 2003, which were different from CLEF 2001 and 2002 collections. So, we made the challenging decision of using only the training topics from CLEF 2003. All the tests were carried out with the same setup (except for language specific resources). For each test, every topic was classified in three categories: "OK" if its average precision was $>MAP$; "bad" if it was only $>MAP/2$, and "hard" if it was $<MAP/2$. In fact, it is not easy to decide when a topic is hard, because the decision depends on the retrieval system, the document collection, and the topic itself. Other authors prefer to use different rules, e.g. a topic is hard if $P10=0$. On the basis of this classification, our aim was to improve hard topics, taking into account that MAP and GMAP of all the topics would be the criteria of robustness in the task. We also had in mind long (title + description) and short (title) queries.

Our main focus was monolingual retrieval. The steps followed are explained below. For bilingual and multilingual retrieval experiments we used machine translation (MT) programs to translate topics into document language, and

then we performed a monolingual retrieval. A weighted max-min normalization method was used for merging lists in multilingual retrieval.

2 Monolingual Retrieval

Our information retrieval system was based on the vector space model. No additional plugins for word sense disambiguation or other linguistic techniques were used. In our experiments, the first step was to obtain a good term-weighting scheme to be used as our basis. The second one was to check whether stopword removal or stemming processes improved the systems robustness. The last step was to apply some local query expansion techniques to see which one (if any) improved the performance of the least effective topics.

Term weighting. In our experiments we converted all words to lowercase, suppressed the stress signs, and included numbers as index terms. A lot of tests were carried out with training topics to obtain the best document-query weighting scheme. The best scheme for all the collections was **dnu-ntc**. For documents, letter *u* stands for the *pivoted document normalization* [4]: we adjusted *pivot* to the average document length, so we had to refine *slope*. Curiously, the best performance was reached for all the collections with $slope = 0.1$.

Stopword removal. Stop words are those which appear very often or which have little meaning. They are usually removed in the indexing process. It is not easy to build a good set of stop words. We decided to remove the most frequent words according to three thresholds, that is, those which appeared in more than 15, 25 and 50 percent of all documents. There was hardly any difference in MAP and GMAP when stop words were removed (for all thresholds). But it is important to note that for the English and Italian collections, removing stop words improved slightly the performance of hard topics. No change was observed for the French and Spanish collections. We decided to remove the terms present in more than 25 percent of documents, which was the best threshold in our tests.

Stemming. Stemming can be conceived as a mechanism for query expansion, since each word is expanded with other words that have the same stem. For each language we used a different stemmer: for English the well-known Porter stemmer, and for French, Italian and Spanish the stemmers from the University of Neuchâtel in the web page <http://www.unine.ch/info/clef/>. Stemming improved MAP and GMAP for all the collections, but hard topics behaved differently. For the English collection, hard topics scarcely changed their performance. For the Spanish collection hard topics had a little improvement. For the French and Italian collections the improvement was noticeable. So we decided to use stemming for all the collections.

Topic difficulty. As in TREC 2004, we tried to find some heuristics to predict topic difficulty without relevance information, and to use it to make decisions at run time. We used term frequency and document frequency of the topic terms, but no heuristics was found, so we discontinued this procedure.

2.1 Local Query Expansion

Blind Relevance Feedback. BRF is a common technique used for local query expansion. The Rocchio formula is the most frequently used for this purpose. A lot of tests were carried out to obtain the best performance for all the collections, using the first 5, 10, 15 and 20 retrieved documents for expansion. The best results were achieved with $\alpha = 1$, $\beta = 2.5$, taking the first 5 retrieved documents, and using an expanded query with about 50 terms.

For the English collection, BRF deteriorated MAP and GMAP, as well as hard topics. For the French, Italian and Spanish collections, BRF improved only a little MAP and GMAP. For these collections, no changes were found in hard topics, except in tests where only the topics title field was used, in whose cases a little improvement was found.

Co-occurrence Based Thesauri. Term co-occurrence has been frequently used in IR to identify some of the semantic relationships that exist among terms. Several coefficients have been used to calculate the degree of relationship between two terms. All of them measure the number of documents in which they occur separately, in comparison with the number of documents in which they co-occur. By computing co-occurrence values for all terms we obtain a co-occurrence matrix. The aim of using this matrix is to expand the entire query, not only separate terms. To expand the original query, terms with a high co-occurrence value with all terms of the query must be selected. We use the measurement of scalar product with all query terms to obtain the terms with highest potential to be added to the original query. A description of this procedure can be found in [5].

To carry out this expansion technique we used two co-occurrence units: (a) the complete document, and (b) windows of terms close to query terms. In the first case only two aspects have to be tuned: the number of retrieved documents used to build the co-occurrence matrix, and the number of expanded terms to be added to the original query. We carried out a lot of tests for this purpose, using three co-occurrence coefficients: Tanimoto (Jaccard), cosine and Dice [6]. The results were discouraging, as the retrieval performance was deteriorated in all the tests. The reason for this was the weighting scheme for topics, **ntc**: the query vector was normalized, and the added terms gained higher weight than the original terms. We repeated the tests using **ntn** scheme, and we obtained better results. The best improvement was achieved with 5 or 10 documents, expanding the original query with about 15 terms, and using the Tanimoto coefficient. All the collections, except for the English one, improved performance with these settings. For hard topics, improvements were observed only for the Spanish and French collections, and only for long queries, i.e. using title and description fields of topics.

VIENNA: **Letter-bomb** attacks yesterday injured two women in the Austrian town of Linz and another woman in Munich, Germany. However they missed one target -- 29-year-old Arabella **Kiesbauer**, a television talk show hostess of mixed race and outspoken human rights campaigner. Police linked *the bombs* to anti-foreigner attacks in Austria by suspected right-wing extremists in the past 18 months. **Letter bombs** strike women.

Fig. 1. Example of terms windows created in a document for topic #141, *Letter Bomb for Kiesbauer*; distance value set to 1

Co-occurrence Based Thesauri Built Using Terms Windows. This technique is identical to the previous one, but it uses windows of terms to obtain the co-occurrence relations instead of the whole document. Terms close to query terms must have higher value of relation than other terms. In this case, it is important to define the distance value between two terms. If distance is zero, both terms are adjacent. If distance is one, then there exists one term between the two terms, and so on. A new window is created when a query term appears in the document: close terms are included in the window depending on distance. If another query term appears inside the window, then the window enlarges its limits to encompass other terms within the same distance. Thus, more terms will co-occur with other query terms (see Fig. 1). To compute the distance, stop words are removed and sentence or paragraph limits are not taken into account.

For all the collections we carried out a lot of tests to obtain the best settings. The highest improvement achieved with this expansion technique was by using a distance value of 1 or 2, taking the first 5 or 10 retrieved documents, and adding about 30 terms to the original query. We only used the Tanimoto co-occurrence coefficient. With these settings, MAP and GMAP improved for all the collections, and in general that was also the case for hard topics (for the Italian collection a little deterioration was found in the tests with short topics).

This expansion technique was the best one for all our tests, and we decided to submit only the relevant runs to the CLEF 2006 Robust Task organization. Table 1 shows the evolution of MAP and GMAP of our monolingual experiments.

3 Bilingual Retrieval (Italian to Spanish)

Our CLIR system was the same as that used in monolingual retrieval. A previous step was carried out before searching, to translate Italian topics into Spanish. We used two MT programs: L&H Power Translation Pro 7.0 and Worldlingo¹. The resulting translations were not post-edited. For each topic we combined the terms of the translations in a single topic: this is another expansion process, although in most cases the two translations were identical. Finally, a monolingual

¹ <http://www.worldlingo.com>

Table 1. Percentage of improvement of MAP and GMAP for our monolingual experiments with local query expansion (using only training topics from CLEF 2003)

		Basis dnu-ntc		BRF				Co-occurrence Thesauri		Terms Windows	
				$\alpha = 1, \beta = 1$		$\alpha = 1, \beta = 2.5$					
		MAP	GMAP	MAP	GMAP	MAP	GMAP	MAP	GMAP	MAP	GMAP
English	td	.4691	.2957	-1.36	-3.99	-1.15	-7.31	-1.56	-7.47	4.14	2.74
	t	.4289	.2173	2.33	0.32	0.72	-11.46	2.15	-24.76	3.64	3.18
French	td	.5209	.3213	0.56	1.40	0.56	11.17	1.88	14.07	0.19	13.13
	t	.4474	.2338	3.02	7.66	4.49	12.75	8.20	6.50	6.30	18.99
Italian	td	.4846	.3111	3.01	2.54	5.20	3.89	4.89	3.25	2.29	0.09
	t	.4177	.2227	3.42	3.77	7.88	7.54	8.47	0.04	8.00	5.07
Spanish	td	.4679	.3500	4.17	3.29	6.73	4.51	8.61	7.49	8.46	6.31
	t	.4145	.3024	5.62	5.89	5.89	5.95	10.74	6.75	9.51	6.08

retrieval was performed. The local query expansion using co-occurrence based thesauri built with terms windows was also applied.

4 Multilingual Retrieval

In the multilingual subtask we use Spanish as topic language and the document collections in English, French, Italian and Spanish. In our experiments, Spanish topics were translated into the language of each document collection using these MT programs:

- English: L&H Power Translator Pro 7.0, Systran² and Reverso³.
- French: Systran and Reverso.
- Italian: L&H Power Translator Pro 7.0 and Worldlingo.

The resulting translations were not post-edited. For every topic and target language we combined the terms of translations to obtain a single topic, and then we performed monolingual retrieval. We also applied the same local query expansion using windows of terms. Before merging the lists of retrieved documents of all target languages, and the one from the monolingual Spanish experiment, it was necessary to do a weighted max-min normalization in each list: $sim_{new} = factor * (sim - min) / (max - min)$. We considered a weighting factor of 1.02 for the Spanish list, and 1 for the rest. The submitted runs were built selecting the first 1,000 documents with the highest similarity value.

5 Results

It should be noted that the runs we submitted to the CLEF 2006 Robust Task organization had an error: for the topics from CLEF 2001 and CLEF 2002 we

² <http://www.systransoft.com>

³ <http://www.reverso.net>

Table 2. Best entries for the robust subtasks in which we participated. We include the correct positions of our runs. The last column shows results for our runs with only the title field of the topics.

Run	Title & Description						Title
Monolingual Subtask							
	Correct	1st	2nd	3rd	4th	5th	
English	reina	hummingbird	reina	dcu	daedalus	colesir	
MAP	.4843	.4763	.4366	.4348	.3969	.3764	.4560
GMAP	.1208	.1169	.1053	.1011	.0893	.0841	.0915
French	reina	unine	hummingbird	reina	dcu	colesir	
MAP	.4901	.4757	.4543	.4458	.4108	.3951	.4869
GMAP	.1632	.1502	.1490	.1432	.1200	.1191	.1608
Italian	reina	hummingbird	reina	dcu	daedalus	colesir	
MAP	.4496	.4194	.3845	.3773	.3511	.3223	.4187
GMAP	.1259	.1147	.1055	.0919	.1050	.0823	.0999
Spanish	reina	hummingbird	reina	dcu	daedalus	colesir	
MAP	.5380	.4566	.4401	.4214	.4040	.4017	.4780
GMAP	.3162	.2361	.2265	.2132	.1964	.1884	.2482
Bilingual Subtask (IT→ES)							
	Correct	1st	2nd	3rd			
Spanish	reina	reina	dcu	daedalus			
MAP	.4411	.3693	.3322	.2689			.3628
GMAP	.1864	.1342	.1044	.0619			.0826
Multilingual Subtask (ES→[EN ES FR IT])							
	1st	Correct	2nd	3rd	4th		
Multilingual	jaen	reina	daedalus	colesir	reina		
MAP	.2785	.2347	.2267	.2263	.1996		.2070
GMAP	.1569	.1672	.1104	.1124	.1325		.1357

included documents that existed only in CLEF 2003 collections, so our runs showed wrong measurements. Table 2 shows the results of mandatory runs (title + description) in all the subtasks where we participated. We include in bold the correct positions of our runs when incorrect documents have been removed. The MAP and GMAP values of correct runs with short queries (only the title field of topics) are shown in the last column.

In monolingual and bilingual retrievals our correct mandatory runs obtain the best ranking in all the subtasks where we participated. In the multilingual subtask, our correct mandatory run is in the upper part of the ranking. For short queries we also obtain high values of MAP and GMAP measurements.

6 Conclusions

Local query expansion using co-occurrence based thesauri built with windows of terms is an effective and simple expansion technique, and can be easily

implemented without adjusting many parameters. It is important to note that for thesauri construction only the information from retrieved documents was used. We also used a simple document retrieval system based on the vector space model, and we looked for a good document-query weighting scheme as the basis for our next expansion experiments. With these settings, our correct mandatory runs obtain the best rankings in the monolingual subtask for all languages in which we participated.

For the bilingual Italian to Spanish subtask, collecting terms from some translations of a topic is another query expansion technique that, combined with the one we used in monolingual experiments, improves the system's performance. Our mandatory run was the best run in the subtask.

For multilingual retrieval we used the same procedure to obtain the retrieved document lists for each target language, and then we combined them through weighted max-min normalization. In this subtask our correct mandatory run also ends up in good ranking.

It is important to note that for short queries we have obtained high values of MAP and GMAP measurements. Thus, the local query expansion technique we applied is also a good expansion technique for short queries.

Acknowledgement

This work was partially supported by a Spanish Science and Education Ministry project grand, number EA2006-0080.

References

1. Voorhees, E.M.: Overview of the TREC 2003 robust retrieval track. In: The Twelfth Text REtrieval Conference, NIST (2003)
2. Voorhees, E.M.: Overview of the TREC 2004 robust retrieval track. In: The Thirteen Text REtrieval Conference, NIST (2004)
3. Voorhees, E.M.: Overview of the TREC 2005 robust retrieval track. In: The Fourteenth Text REtrieval Conference, NIST (2005)
4. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedings of the 19th Annual International ACM SIGIR Conference, pp. 21–29 (1996)
5. Zazo, A.F., Figuerola, C.G., Alonso Berrocal, J.L., Rodríguez, E., Gómez, R.: Experiments in term expansion using thesauri in Spanish. In: Peters, C., Braschler, M., Gonzalo, J. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 301–310. Springer, Heidelberg (2003)
6. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)

Dublin City University at CLEF 2006: Robust Cross Language Track

Adenike M. Lam-Adesina and Gareth J.F. Jones

School of Computing, Dublin City University, Dublin 9, Ireland
{adenike,gjones}@computing.dcu.ie

Abstract. The main focus of the DCU group's participation in the CLEF 2006 Robust Track track was to explore a new method of re-ranking a retrieved document set based on the initial query with a pseudo relevance feedback (PRF) query expansion method. The aim of re-ranking using the initial query is to force the retrieved assumed relevant set to mimic the initial query more closely while not removing the benefits of PRF. Our results show that although our PRF is consistently effective for this task, the application of the current version of our new re-ranking method has little effect on the ranked output.

1 Introduction

This paper describes the DCU experiments for the CLEF 2006 Robust Track. Our official submissions included monolingual runs for English and for Spanish, Italian and French where topics and documents were translated into English, and a bilingual run for Spanish using English topics. Our general approach was to translate non-English documents and topics into English for use as a pivot language using Systran Version: 3.0. Pseudo Relevance Feedback (PRF) using a summary-based approach, shown to be effective in our submissions to previous CLEF workshops, was applied. In addition, for this task we explored the application of a new post-retrieval re-ranking method that we are developing.

The remainder of this paper is structured as follows: Section 2 describes our system setup and the IR methods used, Section 3 presents our experimental results and Section 4 concludes the paper with a discussion of our findings.

2 System Setup

Basic Retrieval System. For our experiments we used the City University research distribution version of the Okapi system retrieval system. Stopwords were removed from both the documents and search topics, and the Okapi implementation of Porter stemming algorithm [1] was applied to both the document and search terms. The Okapi system is based on the BM25 weighting scheme [2]. In our experiments values of the $k1$ and b BM25 parameters were estimated using the CLEF 2003 ad hoc retrieval task data.

Pseudo-Relevance Feedback. Short and imprecise queries can affect information retrieval (IR) effectiveness. To lessen this negative impact, relevance feedback (RF)

via query expansion (QE) is often employed. QE aims to improve initial query statements by addition of terms from user assessed relevant documents. Pseudo-Relevance Feedback (PRF) assuming top ranked documents are relevant, can result in a query drift if expansion terms are selected from assumed relevant document which are in fact not relevant. In our past research work [3] we discovered that although a top-ranked document might not be relevant, it often contains information that is pertinent to the query. Thus, we developed a PRF method that selects appropriate terms from document summaries. These summaries are constructed in such a way that they contain only sentences that are closely related to the initial query. Our QE method selects terms from summaries of the top 5 ranked documents. The summaries are generated using the method described in [3]. For all our experiments we used the top 6 ranked sentences as the summary of each document. From this summary we collected all non-stopwords and ranked them using a slightly modified version of the Robertson selection value (*rsv*) [2]. The top 20 terms were then selected in all our experiments. In our modified version, potential expansion terms are selected from the summaries of the top 5 ranked documents, and ranked using statistics from assuming that the top 20 ranked documents from the initial run are relevant.

2.1 Re-ranking Methodology

As part of our investigation for the CLEF 2006 robust track we explored the application of a novel re-ranking of the retrieved document list obtained from our PRF process. This reordering method attempts to ensure that retrieved documents with more matching query terms have their ranking improved, while not discarding the effect of document weighting scheme used. To this end we devised a document re-ranking formula as follows:

$$\frac{doc\_wgt}{(1-b) + (b*nmt/mmt)} \quad (1)$$

where *doc_wgt* = the original document matching score

b = an empirical value ranging between 0.1 and 0.5

nmt = the number of original topic terms that occur in the document

mmt = mean value of *nmt* for a given query over all retrieved documents.

3 Experimental Results

In this section we describe our parameter selection and present our experimental results for the CLEF 2006 Robust track. Results are given for baseline retrieval without feedback, after the application of our PRF method and after the further application of our re-ranking procedure. Our experiments used the Title and Description (TD) fields or Title, Description and Narrative (TDN) fields of the topics. For all runs we present the precision at both 10 and 30 documents cutoff (P10 and P30), standard TREC average precision results (AvP), the number of relevant documents retrieved out of the total number of relevant in the collection (RelRet), and the change in number of RelRet compared to Baseline runs.

3.1 Selection of System Parameters

To set appropriate system parameters development runs were carried out using the training topics provided. The topics provided were taken from the CLEF 2003 The Okapi parameters were set as follows $k1=1.2$ $b=0.75$. For all our PRF runs, 5 documents were assumed relevant for term selection and document summaries comprised the best scoring 6 sentences in each case. Where the number of sentences was less than 6, half of the total number of sentences were chosen. The *rsv* values to rank the potential expansion terms were estimated based on the top 20 ranked assumed relevant documents. The top 20 ranked expansion terms were added to the original query in each case. Based on results from our previous experiments, the original topic terms are upweighted by a factor of 3.5.

3.2 Experimental Results

Table 1 summarises the results of our experiments. Results are shown for the following runs:

Baseline – baseline results without PRF using TDN topics fields

f20narr – feedback results using the TDN topic fields with 20 terms added

f20re-ranked - same as f20narr, but documents are re-ranked using equation (1)

f20desc – feedback results using the TD topics fields with 20 terms added

Comparing the Baseline and f20narr runs it can be seen that application of PRF improves all the performance measures for all runs with the exception of the RelRet

Table 1. Results for Baseline, PRF and re-ranked runs results for the CLEF 2006 Robust track

	Run-ID	English	French	Spanish	Italian	Spanish bi
Baseline (TDN)	P10	0.422	0.395	0.485	0.382	0.357
	P30	0.265	0.269	0.351	0.262	0.266
	Av.P	0.544	0.470	0.445	0.388	0.314
	RelRet	1496	2065	4468	1736	3702
F20narr (TDN)	P10	0.436	0.425	0.507	0.434	0.413
	P30	0.276	0.294	0.375	0.296	0.300
	Av.P	0.558	0.504	0.478	0.459	0.357
	RelRet	1508	2091	4413	1779	3856
	Chg RelRet	+12	+26	-55	+43	+154
F20re-ranked (TDN)	P10	0.433	0.424	0.509	0.434	0.407
	P30	0.276	0.295	0.377	0.296	0.298
	Av.P	0.558	0.508	0.480	0.459	0.358
	RelRet	1507	2092	4426	1783	3900
	Chg RelRet	+11	+27	-42	+47	+198
F20desc (TD)	P10	0.396	0.370	0.450	0.398	0.386
	P30	0.261	0.272	0.358	0.279	0.288
	Avep	0.494	0.452	0.435	0.419	0.343
	RelRet	1493	2074	4474	1778	3759
	Chg RelRet	+3	+9	+6	+42	+57

for Spanish monolingual where there is a small reduction. By contrast for the Spanish bilingual run there is a much larger improvement in RelRet than is observed for any of the other runs. Application of the re-ranking method to the f20narr list produces little change in the ranked output. The only notable change is a further improvement in the RelRet for the Spanish bilingual task. Varying the value of the b factor in equation 1 made only a small difference to the results. We are currently investigating the reasons for these results, and exploring approaches to the re-ranking method which will have a greater impact on the output ranked lists.

The PRF using only the TD topic fields show that the addition of the N field produces an improvement in retrieval effectiveness in call cases.

4 Conclusions

This paper has presented a summary of our results for the CLEF 2006 Robust Track. The results show that our summary-based PRF method is consistently effective across this topic set. We also explored the use of a novel post-retrieval re-ranking method. Application of this procedure led to very little modification in the ranked lists, and we are currently exploring alternative variations on this method.

References

- 1 Porter, M.F.: An algorithm for suffix stripping. *Program* 14, 10–137 (1980)
- 2 Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In: Harman, D.K. (ed.) *Proceedings of the Third Text REtrieval Conference (TREC-3)*, NIST, pp. 109–126 (1995)
- 3 Lam-Adesina, A.M., Jones, G.J.F.: Applying Summarization Techniques for Term Selection in Relevance Feedback. In: *Proceedings of the ACM SIGIR 2001*, New Orleans, pp. 1–9. ACM, New York (2001)

JHU/APL Ad Hoc Experiments at CLEF 2006

Paul McNamee

Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099 USA
paul.mcnamee@jhuapl.edu

Abstract. JHU/APL continued its participation in multilingual retrieval at CLEF in 2006. We again applied our hallmark technique for combating language diversity and morphological complexity: character n -gram tokenization. This year we participated in the ad hoc cross-language track and submitted both monolingual and bilingual runs. Our experimental results this year agree with our previous reports that n -grams perform especially well in linguistically complex languages, notably Bulgarian and Hungarian, where monolingual improvements of 27% and 70% respectively were observed compared to space-delimited word forms. As in CLEF 2005, our bilingual submissions made use of subword translation, statistical translation of character n -grams using aligned corpora, when parallel data were available, and web-based machine translation, when no suitable data was available to us.

Keywords: Cross-language information retrieval, character n -gram tokenization, corpus-based translation.

1 Overview

This was the seventh year that the Johns Hopkins University Applied Physics Laboratory (JHU/APL) participated in the Cross-Language Evaluation Forum workshop. Due to limited time and resources we were only able to participate in the ad-hoc task for which we submitted both monolingual and bilingual runs. For our experiments in CLEF 2006 we again used the HAIRCUT retrieval system, which notably supports character n -gram tokenization (sequences of n consecutive characters), which has been a consistent theme in JHU/APL's information retrieval research. Our experiments this year are patterned after the same, fairly successful approach taken for CLEF 2005, which we briefly describe below.

Because we endeavor to keep language-specific processing to a minimum we treat each language practically identically. We remove case information and diacritical marks, consider whitespace to denote word boundaries, and generate overlapping character n -grams from the stream of words. In some experiments we perform word-based indexing or use stemmed forms produced from the Snowball stemmer [8]. We make no attempt at segmentation or decompounding in any of the CLEF languages. For our n -grams we usually select fixed lengths of $n=4$ or $n=5$, which we have established as the best lengths to optimize retrieval performance [3]. Regardless of which form of tokenization we employ (*i.e.*, words, stems, or n -grams) we use a

statistical language model method to estimate document relevance [1]. To produce our ranked document list we perform the following processing stages:

- remove stopterms and query phrases (e.g., find documents that);
- [bilingual] apply pretranslation query expansion using comparable collections (i.e., the source language CLEF subcollection, when available);
- [bilingual] statistical translation using aligned parallel texts – from a 500MB/language collection obtained from the *Official Journal of the European Union (1998-2004)* – or other translation such as available online machine translation;
- psuedo-relevance feedback (queries are expanded to 60 terms; terms are weighted using 20 top-ranked and 75 low-ranked documents from the top 1000);
- and, merging of multiple independent runs to produce a single consensus run.

We emphasize that our non-MT statistical translation is based on single-term projection from a source language to the target language and may involve translation of n-grams which we call *subword translation* [4]. Subword translation attempts to overcome obstacles in dictionary-based translation, such as word lemmatization, matching of multiple word expressions, and inability to handle out-of-vocabulary words such as common surnames [6].

For greater detail about our approach to document and query processing see our report for CLEF 2005 [5].

2 Monolingual Task

For our monolingual work we created indexes for each language using the permissible document fields appropriate to each collection. Our four basic methods for tokenization were unnormalized words, stemmed words obtained through the use of the Snowball stemmer (in English and French), 4-grams, and 5-grams. Information about each index is shown in Table 1 (below). The Snowball stemmer does not have modules to handle Bulgarian or Hungarian.

Table 1. Summary information about the test collection and index data structures

language	#docs	#rel	index size (MB) / unique terms (1000s)			
			words	stems	4-grams	5-grams
BG	67341	778	57 / 67	---	154 / 193	251 / 769
EN	166754	2063	143 / 302	123 / 236	504 / 166	827 / 916
FR	177450	2537	129 / 328	107 / 226	393 / 159	628 / 838
HU	49530	939	59 / 549	---	121 / 150	200 / 741
PT	210734	2904	178 / 418	140 / 254	529 / 174	868 / 907

We observed effects due to tokenization this year that were consistent with our experiments in CLEF 2005. In the Bulgarian and Hungarian languages, substantial benefits were seen when n-grams were used instead of words. We saw 27% and 70%

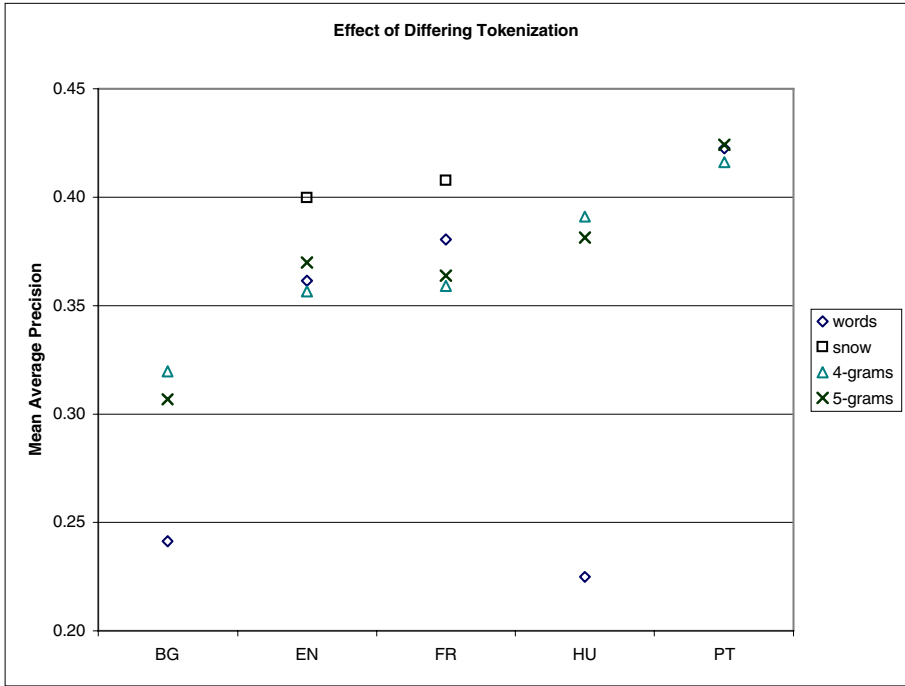


Fig. 1. Relative effectiveness of tokenization methods on the CLEF 2006 test sets

relative improvements, respectively (compared to 30% and 63% in 2005). In the other languages, n-grams performed similarly to words and somewhat worse than the use of stemmed words in English and French. The effect of tokenization on mean average precision is shown Figure 1.

In Bulgarian and Hungarian it seems that 4-grams may have a slight advantage over 5-grams, though testing should be performed to verify that the differences are statistically significant. However, the use of n-grams over unlemmatized words seems strongly indicated.

Our monolingual submissions were either base runs (*i.e.*, those using a single form of tokenization) or a combination of several base runs using various options for tokenization. Our method for combination is to normalize scores by probability mass and to then merge documents by score. All of our submitted runs were automatic runs and used the title and description topic fields and in some cases, the narrative section. Our run names are created by concatenating the query year (02 or 95), the token *apl*, the token *mo* (for monolingual), a two character language code (*e.g.*, *bg* or *fr*), *td* or *tdn* to designate the topic fields used, and any combination of the symbols '4', '5', 'w', or 's' to identify the tokenization strategy. Monolingual performance based on mean average precision is reported in Table 2.

Table 2. Official results for monolingual task

	run id	Fields	Terms	MAP	Rel. Found	Relevant
BG	02aplmobgtd4	TD	4-grams	0.3198	1156	1249
	02aplmobgtdn4	TDN	4-grams	0.3148	1108	1249
FR	95aplmofrtd5s	TD	5-grams, stems	0.4096	1892	2148
	95aplmofrtd4	TD	4-grams	0.3592	1503	2148
	95aplmofrtdn5s	TDN	5-grams, stems	0.4290	1884	2148
HU	02aplmohutd4	TD	4-grams	0.3911	1267	1308
	02aplmohutdn4	TDN	4-grams	0.3943	1272	1308
PT	95aplmopttd4	TD	4-grams	0.4162	2292	2677
	95aplmopttd5	TD	5-grams	0.4242	2312	2677
	95aplmopttdn5	TDN	5-grams	0.4321	2378	2677

3 Bilingual Task

Our preferred approach to bilingual retrieval adds the following two steps to our normal monolingual processing: (1) apply pre-translation query expansion using the source language CLEF corpus and (2) translate terms statistically using aligned parallel corpora. We have had good success using aligned parallel corpora to extract statistical translations. Others have also relied on corpus-based translation; however, we recently demonstrated significant improvements in bilingual performance by translating character n-grams directly. We call this ‘*subword translation*’. Additionally we also translate stemmed words and words. This is the approach we used for our X->French and X->Portuguese runs.

Our primary translation resource is a parallel corpus derived from the Official Journal of the E.U. [7], which is published in the official languages. We started with 80 GB of PDF files and extracted approximately 750 MB of alignable plain text per language. We are aware of other sizeable multi-aligned corpora (*e.g.*, Philip Koehn’s *Europarl* corpus [2]) but do not use additional data for reasons of simplicity.

Though pre-translation query expansion can be used to produce any term type (*e.g.*, n-grams or stems) we have been extracting 60 unnormalized words from source language documents and then translating terms (words, stems, or n-grams) derived from these words. Our method for inducing translations is to take a source language term and identify documents in the parallel data containing the term and to dynamically compute term frequency information for cross-lingual collocations. Each target language term is given a weight using term similarity statistics (*i.e.*, chi-squared, mutual information). Because we have previously observed good performance by doing so, we select just the single highest weighted target term to use in the final query. This method for producing candidate translations does not require translation model software such as GIZA++.

To support translation for language pairs that the *Official Journal* data did not include, we relied on query translation using web-based machine translation. We used MT to translate from Bulgarian and Indonesian for queries against English documents. The online services we used are located at:

- http://www.toggetext.com/kataku_trial.php (IN to EN)
- http://www.bultra.com/test_e.htm (BG to/from EN)

We erroneously ran the wrong year's set of queries for our Bulgarian to English runs, and we were not able to remedy this problem at the last minute; thus we submitted official runs using only Indonesian to English with MT software. Our official bilingual results are shown below in Tables 3 and 4 below.

Table 3. JHU/APL's official results for bilingual task using corpus-based translation

run id	Source	Target	Fields	Terms	MAP	% mono	Rel. Found	Relevant
aplbienfrd	EN	FR	TD	5-grams	0.3360	82.03%	1674	2148
aplbienfrn	EN	FR	TDN	5-grams	0.3536	82.42%	1685	2148
aplbidefrd	DE	FR	TD	5-grams	0.3189	77.86%	1620	2148
aplbidefrn	DE	FR	TDN	5-grams	0.3400	79.25%	1653	2148
aplbienptd	EN	PT	TD	5-grams	0.3529	83.19%	2174	2677
aplbienptn	EN	PT	TDN	5-grams	0.3517	81.39%	2169	2677
aplbiesptd	ES	PT	TD	5-grams	0.3549	83.66%	2131	2677
aplbiesptn	ES	PT	TDN	5-grams	0.3576	82.76%	2140	2677

Table 4. JHU/APL's official results for bilingual task using machine translation

run id	Source	Target	Fields	Terms	MAP	% mono	Rel. Found	Relevant
aplbiiinen4	ID	EN	TD	4-grams	0.3152	88.39%	960	1258
aplbiiinen5	ID	EN	TD	5-grams	0.3257	88.05%	946	1258

In Table 3, our results using corpus-based subword translation achieved bilingual performance between 78% and 84% of our best monolingual runs for the given target language. Table 4 details our results using available machine translation software. Our past experience with MT-based CLIR tells us that performance depends heavily on the individual translation engine used. In this case, using the *Kataku* translator we obtained about 88% of our best monolingual baselines and 4-gram or 5-gram tokenization worked equally as well.

4 Summary

JHU/APL participated in the CLEF 2006 evaluation, submitting runs only for the monolingual and bilingual ad hoc tasks. As in the past we made use of character n-gram tokenization, a language-neutral approach that again served us well this year. In Bulgarian retrieval use of n-grams produced a 27% relative improvement in performance compared to the use of unnormalized words. With Hungarian text, n-grams were markedly superior to words with a 70% relative advantage.

For bilingual retrieval we employed subword translation in several official runs, with good effect. However we lacked readily available parallel corpora for Bulgarian

and Hungarian and we only submitted bilingual runs for Western European languages and using machine translation, from Indonesian to English. As part of translation (with parallel text) or after it (using MT) we used n-grams in cross-lingual retrieval and they worked well, achieving 80-90% of the performance of our best monolingual baseline runs.

References

- 1 Hiemstra, D.: Using Language Models for Information Retrieval. Ph. D. Thesis, Center for Telematics and Information Technology, The Netherlands (2000)
- 2 Koehn, P.: Europarl: A multilingual corpus for evaluation of machine translation. Unpublished, <http://www.isi.edu/~koehn/publications/europarl/>
- 3 McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7(1-2), 73–97 (2004)
- 4 McNamee, P., Mayfield, J.: Translating Pieces of Words. In: *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval (SIGIR-2005)*, Salvador, Brazil, pp. 643–644 (August 2005)
- 5 McNamee, P.: Exploring New Languages at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
- 6 Pirkola, A., Hedlund, T., Keskusalo, H., Järvelin, K.: Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval* 4, 209–230 (2001)
- 7 <http://europa.eu.int/>
- 8 <http://snowball.tartarus.org/>

The Domain-Specific Track at CLEF 2006: Overview of Approaches, Results and Assessment

Maximilian Stempfhuber and Stefan Baerisch

GESIS – Informationszentrum Sozialwissenschaften (IZ), Lennéstraße 30,
53113 Bonn, Germany
{max.stempfhuber, stefan.baerisch}@gesis.org

Abstract. The CLEF domain-specific track uses databases in different languages from the social science domain as the basis for retrieval of documents relevant to a user's query. Predefined topics simulate the user's information needs and are used by the research groups participating in this track to generate actual queries. The documents are structured into individual elements so that they can be used for optimizing the search strategies. One type of the retrieval task was on cross-language retrieval, the finding of information using queries in a different language than the actual language of the documents. English, German and Russian were used as languages for queries and documents. Queries therefore had to be translated from one of the languages to one or more of the other languages. Besides the cross-language tasks also monolingual retrieval tasks were carried out where query and document are in the same language. The focus here was to map the query onto the language and internal structure of the documents.

This paper gives an overview of the domain-specific track and reports on noteworthy trends in approaches and results as well as on the topic creation and assessment process.

1 Introduction

The CLEF domain-specific task aims on cross-language retrieval in English, German and Russian. Queries as well as documents are taken from the social science domain and are provided in all three languages. This section presents the domain-specific task with a focus on the corpora used. It gives an overview of the activities in the CLEF 2006 campaign. The following sections are on the methods used and results achieved in the retrieval, and on the processes of topic-creation and assessment. The paper concludes with an outlook on the future development of the domain-specific track.

2 The CLEF 2006 Domain-Specific Track

The domain-specific track can be divided into 3 subtasks depending on the language of the query and the documents.

- In the monolingual subtask queries and documents are in the same language. The Russian subtask uses the INION Collection while the English und German subtasks work on the GIRT corpus.
- In the bilingual subtasks queries in one language are used with a collection in a different language. The domain-specific track includes 6 bilingual subtasks for all combinations of English, German, and Russian.
- The multilingual task uses queries in one language against the combination of the content of all collections.

The query tasks in CLEF are provided to the participating groups as topics. Each topic is the representation of an information need from the social science domain. Topics as well as documents are structured. A topic includes a title, a short description of the information need and a longer narrative for further clarification. Documents, among other information, include fields such as author, title and year of publication, terms from a thesaurus and an abstract:

Table 1. Sample Topic

Title	Childlessness in Germany
Description	Find information on the factors for childlessness in Germany.
Narrative	All documents examining causes and consequences of childlessness among couples in Germany are of interest. This also includes documents covering the topic's macroeconomic aspects.

The participating groups submit one or more runs for each subtask. A run includes the 1000 highest ranked documents for each of the 25 topics. The runs from all groups for a given subtask are pooled, the highest ranking documents are then assessed by domain experts in order to obtain a reference on the relevance of the documents.

2.1 The GIRT4 and INION Collections

GIRT, the German Indexing and Retrieval Testdatabase is a pseudo-parallel corpus with 302.638 documents in English and German. The documents cover primarily the domain of the German Social Sciences, in particular scientific publications. Details about the Girt corpus, especially about the development of the current fourth version, can be found in [1,2]. The INION corpus covers the Russian social sciences and economics. The collection includes 145.802 documents and replaces the Russian Social Science Corpus used in 2005.

2.2 Participants and Submitted Runs

The University of California in Berkeley, the University of Hagen in Germany, the University of Technology in Chemnitz and the University of Neuchatel took part in the 2006 domain specific track, submitting 36 runs in total: 12 bilingual runs and 2 multilingual runs. For detailed figures see the following table:

Table 2. Submitted runs

Sub-task	# Participants	# Runs	Topic Language
Multi-lingual	1	2	DE:1 EN:1
Bilingual X → DE	2	6	EN:6
Bilingual X → EN	1	3	DE:2 RU:2
Bilingual X → RU	1	3	EN:2 DE:1
Monolingual DE	4	13	
Monolingual EN	3	8	
Monolingual RU	1	1	
Total	4	36	

2.3 Methods and Results Overview

Details on the results and employed methods of the groups participating in this track are given in the corresponding papers of this volume. This chapter gives only a brief overview of the methods employed. The University of Hagen used a re-ranking approach based on antagonistic terms, improving the mean average precision. The Chemnitz Technical University used Apache Lucene as the basis for the retrieval and achieved the best results with a combination of suffix stripping, stemming, and de-compounding. The University of California, Berkeley, did not use methods for thesaurus-based query expansion and de-compounding as they were employed before, because they wanted to establish a new baseline for future research. The University of Neuchatel used DFR GL2 and Okapi with several combinations of the fields included in a topic against the different fields of GIRT4 documents.

2.4 Topics

The team responsible for the topic-creation process changed in 2006. This led to changes in the topic-creation process itself as well. In order to produce a large number of potential information needs covering a wide area of the social sciences, a number of domain experts familiar with the GIRT corpus were asked for contributions to the topics. 25 topics were selected from 42 submissions. These topics were compared to the topics from the years 2001-2005 in order to remove topics too similar to topics already used. Removed topics were replaced from the pool of previously unused topics. Involving domain experts with detailed knowledge about the corpus resulted in many topics of interest but also caused more effort in the selection of the actual set of topics. This was due to the fact that not all submitters of topics had detailed knowledge of the CLEF rules for topic creation and corrections had to be made afterwards to provide consistent structure and wording. For future CLEF-campaigns, this process will be further optimized.

The definition of topics suitable for the GIRT corpus as well as the Russian INION corpus remains challenging. Lack of Russian language skills in the topic-creation team increases the difficulty of creating a set of topics well suited for both corpora. An additional factor in 2006 was lack of experience with the INION corpus. Also in this regard, we hope to improve the topic creation process in the next campaign as the

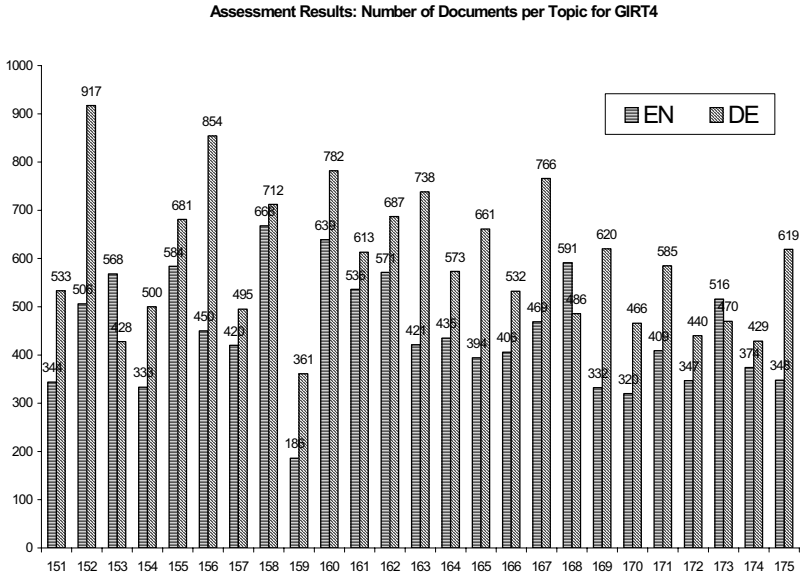


Fig. 1a. Number of documents per topic

multi-lingual corpus is now stable again. Nevertheless, there was no negative effect of the changes in topic creation on the results of the participating groups.

2.5 Assessment Process

As in 2005, the assessment for English and German were done using a Java Swing program while our partner from the Research Computing Center of the M.V. Lomonosov Moscow State University used the CLEF assessment tool for the assessment of the Russian documents. The decision to use this software for English and German was made based on the fact that the assessments of English and German documents from the pseudo-parallel corpus should be compared to each other. An interesting finding was the statement of both assessors that they liked the ability to choose between keyboard-shortcuts and a mouse driven interface since it allowed for different working styles in the otherwise monotonous assessment task. Figures 1a and 1b show the number of documents in the pool for each topic and the percentage of relevant documents. It should be noted that, while the number of documents found for each topic is relatively stable, the percentage of relevant documents (see figure 1b) shows outliers with the topics number 167 and 169. Investigation on this outliers and discussion with the assessors showed that the outliers were caused by complex exclusion criteria in case of topic 167 and the use of the complex concepts of gender specific education and the specifics of the German primary education system in case of topic 169.

The reassessment of documents present in both the English and the German pools of the parallel corpus – first introduced in the 2005 campaign – was repeated in 2006. When such a document had different relevance judgements between the two language pools, a second round of assessment followed a discussion between the assessors on their individual criteria for the judgements in question.

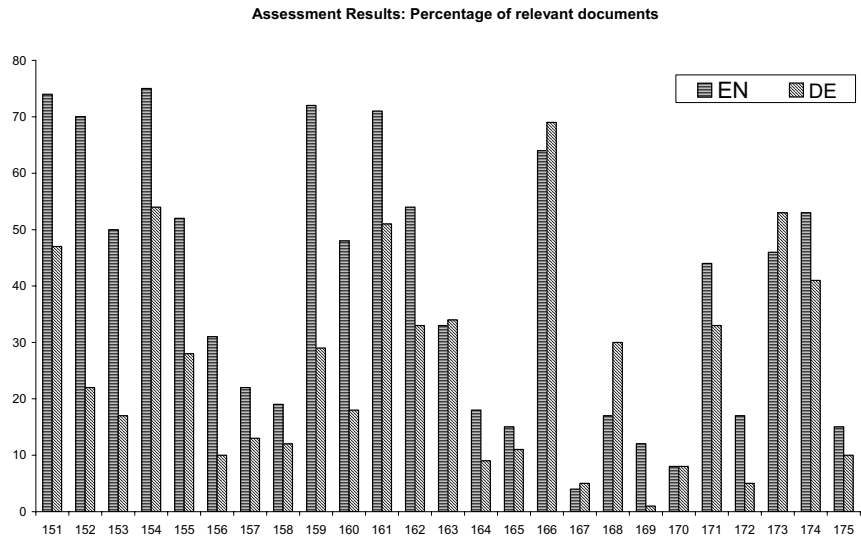


Fig. 1b. Percentage of relevant documents per topic

In addition to the reassessment of pairs a second round of assessment on the topic-level was conducted in 2006. This reassessment focused on the topics with the most significant difference in the percentage of relevant documents between German and English.

3 Results

The detailed information on recall and precision for the participants and their runs can be found in the groups chapters in this volume [2,3,4,5] and shall not be repeated here

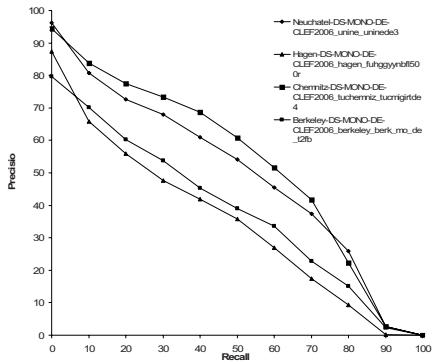


Fig. 2. DS Monolingual Task German

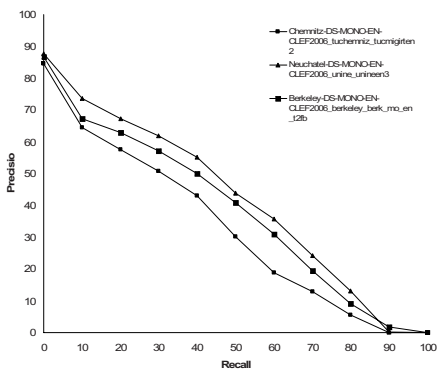


Fig. 3. DS Monolingual Task English

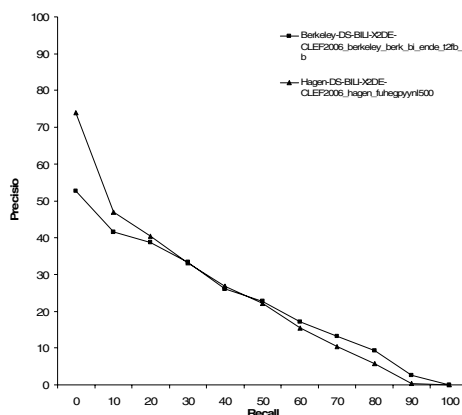


Fig. 4. DS Bilingual Task English to German

in full. The Recall-Precision graph for the tracks which attracted the most participants can be found in the figures 2, 3 and 4.

4 Outlook

Several methods are being investigated to improve the quality of the topic creation and the assessment process. In order to create topics covering a broader scope of the social sciences as well as realistic information needs it is planned to involve more domain experts in the topic creation process in the future. These experts will be asked for topic proposals in the form of narratives. The most suitable topics from the resulting pool will then be selected and produced as topics conforming to the topic creation guidelines. In order to better verify the topics against the INION and GIRT4 corpora, a new retrieval system will be set up which contains all corpora and will also allow our Russian partners from the Research Computing Center of the M.V. Lomonosov Moscow State University to better participate in the topic creation process.

Concerning the assessment process several methods are evaluated in order to allow the assessors to better coordinate their assessment criteria. Proposed methods include breaking the assessment und reassessment processes into smaller batches of 5 topics or the notification of the assessors in case of notable differences in the results for a topic.

Finally, to attract more groups to the domain-specific track in 2007, we will try to negotiate an additional native English corpus. We are targeting at the Sociological Abstracts reference database from Cambridge Scientific Abstracts (CSA) which perfectly matches the social science corpora already available in the domain-specific track. For this corpus we are also able to provide bi-directional mappings between the thesaurus used in the GIRT corpus and the one used with the CSA database. This would give the participating groups a broad range of possibilities for query translation between to languages involved in this track.

Acknowledgements

We greatly acknowledge the partial funding by the European Commission through the DELOS Network of Excellence on Digital Libraries. Natalia Lookachevitch from the Research Computing Center of the M.V. Lomonosov Moscow State University helped in acquiring the Russian corpus and organized the translation of topics into Russian and the assessment process for the Russian corpus. Claudia Henning and Jeof Spiro did the German and English assessments.

References

- 1 Kluck, M.: The GIRT Data in the Evaluation of CLIR Systems – from 1997 until 2003. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 379–393. Springer, Heidelberg (2004)
- 2 Kluck, M.: The Domain-Specific track in CLEF 2004: Overview of the Results and Remarks on the Assessment Process. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 260–270. Springer, Heidelberg (2005)
- 3 Kürsten, J., Eibl, M.: Monolingual Retrieval Experiments with a Domain Specific Document Corpus at the Chemnitz Technical University Working Notes for the CLEF 2006 Workshop, September 20-22, Alicante, Spain (2006)
- 4 Larson, R.R.: Domain Specific Retrieval: Back to Basics Working Notes for the CLEF 2006 Workshop, September 20-22, Alicante, Spain (2006)
- 5 Leveling, J.: University of Hagen at CLEF 2006: Reranking documents for the domain-specific task. In: Working Notes for the CLEF 2006 Workshop, September 20-22, Alicante, Spain (2006)
- 6 Savoy, J., Abdou, S.: UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In: Working Notes for the CLEF 2006 Workshop, September 20-22, Alicante, Spain (2006)

Reranking Documents with Antagonistic Terms

Johannes Leveling

FernUniversität in Hagen (University of Hagen)
Intelligent Information and Communication Systems (IICS)
58084 Hagen, Germany
`firstname.lastname@fernuni-hagen.de`

Abstract. For the participation of the IICS group at the domain-specific task (GIRT) of the CLEF campaign 2006, we employ cooccurrence of search terms and antagonistic terms (antonyms or cohyponyms) in documents to derive values for reranking an initial result set.

A reranking test on GIRT 2004 data showed a significant increase in mean average precision (MAP), i. e. a change from 0.2446 MAP to 0.2986 MAP. Precision for the submitted runs for the domain-specific (DS) task did not change significantly, but the setup for the best experiment included a reranking of result documents (0.3539 MAP). For reranking a result set with an already high MAP (provided by the Berkeley group), a significant decrease in precision was observed (MAP dropped from 0.4343 to 0.3653).

1 Introduction

For our participation at the domain-specific information retrieval (IR) task at CLEF 2006, we investigated a method for reranking documents in the initial result set to increase precision.¹ The method determines a set of antagonistic terms, i. e. antonyms or cohyponyms of search terms, and it reduces the score (and subsequently the rank) of documents containing a search term and an antagonistic term. As some terms will frequently occur together with their corresponding antagonistic terms (e. g., *day and night*, *black and white*), cooccurrence of search terms and antagonistic terms is considered as well to adapt the calculation of new scores.

The setup for our previous participations at the domain-specific task was used (see [1]) to obtain initial result sets. It includes of a deep linguistic analysis, query expansion with semantically related terms, blind feedback, an entry vocabulary module [2], and several retrieval functions. For the bilingual experiments, a single online translation service, Prompt², was employed to translate English topic titles and topic descriptions into German.

¹ The research described was in part funded by the DFG (Deutsche Forschungsgemeinschaft) in the project IRSAW (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web).

² <http://www.prompt.ru/>, <http://www.e-prompt.com/>

2 Reranking with Information About Antagonistic Terms

There has already been some research on reranking documents to increase precision in IR. Gey et al. [3] describe experiments with Boolean filters and negative terms for TREC data. In general, this method does not provide a significant improvement, but an analysis for specific topics shows a potential for performance increase. Our group regards Boolean filters to be too restrictive to help improve precision. Furthermore, the case of a cooccurrence of term and filter term (or antagonistic term) in queries or documents is not addressed.

Kamps [4] describes a method for reranking using a dimensional global analysis of data. The evaluation of this method is based on GIRT (German Indexing and Retrieval Testdatabase) and Amaryllis data. The performance gain is lower but on the same order as the increase in precision observed in our first reranking test. While this approach is promising, it relies on a controlled vocabulary and therefore will not be portable between domains or even different text corpora.

We introduce the notion of antagonistic terms, meaning terms with a semantics opposed to the semantics of search terms. For a given search term t , the set of antagonistic terms contains antonyms of t , antonyms of a member in the set of synonyms of t , antonyms of hyponyms of t , antonyms of hypernyms of t , and cohyponyms of t . For example, *animal* and *plant* are antonyms, *vertebrate* and *plant* are antonyms of a hypernym/hyponym, and *vertebrate* and *invertebrate* are cohyponyms. Different semantic information resources were combined to create a semantic net holding this background knowledge, including the computer lexicon HaGenLex [5], a mapping of HaGenLex concepts to GermaNet synonym sets [6], the GIRT-Thesaurus (for hyponym relations) and semantic subordination relations semi-automatically extracted from German text corpora.

Based on queries and relevance assessments from the GIRT task in 2005, cooccurrence of query terms and their antagonistic terms in documents assessed as relevant and in other (non-relevant) documents was calculated. The statistics serve to determine to what amount the score of a document in the result set will be adjusted. For example, a document D with a score S_D that contains a search term A as well as its antonym B will have its score decreased. The reranking algorithm, reranking formula, and statistics on cooccurrence of terms are described in more detail in [7].

3 Reranking Results

The reranking algorithm was tested on data and queries from GIRT 2004 (including the corresponding relevance assessments). Our best official run at the DS task in 2004 has 0.2446 MAP. For different values of a factor c in the reranking formula (see [7]), reranking the results always yielded a significantly higher MAP than the official run: 0.2951 ($c=0.005$), 0.2986 ($c=0.008$), 0.2986 ($c=0.01$), 0.2976 ($c=0.05$), and 0.2759 ($c=0.1$).

For the domain-specific task at CLEF 2006, settings for the following parameters were varied: LA: obtain search terms by a linguistic analysis [1], IRM: select

Table 1. Results for monolingual German and bilingual (German-English) DS experiments

Run Identifier	Parameters			Results	
	Language	LA	IRM	RS	rel_ret MAP
FUHgggydbfl102	DE-DE	Y	<i>tf-idf</i>	N	2444 0.2777
FUHgggydbfl500	DE-DE	Y	BM25	N	2780 0.3205
FUHgggydbfl500R	DE-DE	Y	BM25	Y	2780 0.3179
FUHgggyynbfl500	DE-DE	N	BM25	N	2920 0.3525
FUHgggyynbfl500R	DE-DE	N	BM25	Y	2920 0.3539
FUHegpyynl102	EN-DE	Y	<i>tf-idf</i>	N	2134 0.1980
FUHegpyydl102	EN-DE	Y	<i>tf-idf</i>	N	2129 0.1768
FUHegpyydl500	EN-DE	Y	BM25	N	2422 0.2190
FUHegpyydl500R	EN-DE	Y	BM25	Y	2422 0.2180
FUHegpyynl500	EN-DE	N	BM25	N	2569 0.2448

retrieval model (*tf-idf* or OKAPI/BM25, [8]), and RS: rerank result set ($c=0.025$). Table 1 shows parameter settings and results for DS experiments. Results include the number of relevant and retrieved documents (rel_ret) and MAP.

An additional experiment with an even higher MAP of the initial result set, used results of an unofficial run from the Berkeley group.³ The experiments of the Berkeley group were based on the setup for their participation at the GIRT task in 2005 [9]. Reranking applied on the results set found by the Berkeley group, which has an average MAP of 0.4343 for the monolingual German task, significantly lowered performance to 0.3653 MAP.

4 A Brief Discussion of Results

Reranking experiments for the DS task at CLEF led to different results: 1.) A test based on data and results from CLEF 2004 showed an increase in MAP from 0.2446 to 0.2976 (+ 21.6%) change). 2.) Reranking for the official experiments did not significantly change the MAP. For the monolingual run, MAP dropped from 0.3205 to 0.3179 in one pair of experiments and rose from 0.3525 to 0.3539 in another. For a bilingual pair of comparable experiments (English-German), MAP dropped from 0.2190 to 0.2180. 3.) An additional reranking experiment was based on data provided by the Berkeley group. The MAP decreased from 0.4343 to 0.3653 when reranking was applied to this data.

There are several explanations as to why precision is affected differently. The reranking factor c was not fine-tuned for the retrieval method employed to obtain the initial result set: for the experiments in 2004, we used a database management system with a *tf-idf* IR model, while for GIRT 2006, the OKAPI/BM25 IR model was applied. The corresponding result sets show a different range and

³ Thanks to Vivien Petras at UC Berkeley for providing the data.

distribution of document scores. Thus, the effect of reranking document with the proposed method may depend on the retrieval method employed to obtain the initial results. Furthermore, reranking will obviously become harder the better the initial precision already is. The results from the Berkeley group will be more difficult to improve, as they already have a high MAP.

5 Conclusion

A novel method to rerank documents was briefly presented. It is based on a combination of information about antagonistic relations between terms in queries and documents and their cooccurrence.

For a test with data from CLEF 2004, a performance increase in precision was observed. Official results for CLEF 2006 show no major changes, and an additional experiment based on data from the Berkeley group even shows a decrease in precision. These results indicate that the reranking method has to be fine-tuned to characteristics of an IR system or model.

References

1. Leveling, J.: A baseline for NLP in domain-specific information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 222–224. Springer, Heidelberg (2006)
2. Gey, F.C., Buckland, M., Chen, A., Larson, R.R.: Entry vocabulary – a technology to enhance digital search. In: Proceedings of the First International Conference on HLT (March 2001)
3. Gey, F.C., Chen, A., He, J., Xu, L., Meggs, J.: Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms at TREC-5. In: Proceedings of TREC-5, pp. 181–190 (1996)
4. Kamps, J.: Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In: McDonald, S., Tait, J. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 283–295. Springer, Heidelberg (2004)
5. Hartrumpf, S., Helbig, H., Osswald, R.: The semantically based computer lexicon HaGenLex - Structure and technological environment. *Traitement automatique des langues* 44(2), 81–105 (2003)
6. Baayen, R.H., Piepenbrock, R., Gulikers, L.: The CELEX Lexical Database. Release 2 (CD-ROM). LDC, University of Pennsylvania, Philadelphia, Pennsylvania (1995)
7. Leveling, J.: University of Hagen at CLEF 2006: Reranking documents for the domain-specific task. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop, Alicante, Spain (2006)
8. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Harman, D. (ed.) Proceedings of TREC-3, pp. 109–126 (1994)
9. Petras, V., Gey, F.C., Larson, R.: Domain-specific CLIR of English, German and Russian using fusion and subject metadata for query expansion. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 226–237. Springer, Heidelberg (2006)

Domain Specific Retrieval: Back to Basics

Ray R. Larson

School of Information
University of California, Berkeley, USA
`ray@sims.berkeley.edu`

Abstract. In this paper describe Berkeley's approach to the Domain Specific (DS) track for CLEF 2006. This year we did *not* use the tools for thesaurus-based query expansion and de-compounding for German used in previous years. This year Berkeley submitted 12 runs, including one for each subtask of the DS track. These include 3 Monolingual runs for English, German, and Russian, 7 Bilingual runs (3 X2EN, 1 X2DE, and 3 X2RU), and 2 Multilingual runs. For many DS sub-tasks our runs were the best performing runs, but sadly they were also the *only* runs for a number of sub-tasks. In the sub-tasks where there were other entries, our relative performance was above the mean performance in 2 sub-tasks and just below the mean in another.

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the Domain Specific track. Our submitted runs this year are intended to establish a new baseline for comparison in future Domain Specific evaluations for the Cheshire system, and did not use the techniques of thesaurus-based query expansion or German decompounding used in previous years. This year we used only basic probabilistic retrieval methods for DS tasks. We hope, in future years, (assuming that the task will continue in future years) to be able to use those, or refinements of those techniques via the Cheshire II or Cheshire3 systems.

This year Berkeley submitted 12 runs, including one for each subtask of the DS track. These include 1 Monolingual run for each of English, German, and Russian for a total of 3 Monolingual runs, and 7 Bilingual runs (3 X2EN, 1 X2DE, and 3 X2RU), and 2 Multilingual runs.

This paper first very briefly describes the retrieval methods used, including our blind feedback method for text, which are discussed in greater detail in our ImageCLEF paper. We then describe our submissions for the various DS sub-tasks and the results obtained. Finally we present conclusions and discussion of future approaches to this track.

2 The Retrieval Algorithms

We have discussed the basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions for the Domain Specific track in our

paper for the GeoCLEF track in this volume. In addition we used blind feedback for this track that took the top-ranked 16 term from the top-ranked 14 documents for inclusion in an expanded query. More information about the blind feedback method is available in our GeoCLEF paper in this volume.

3 Approaches for Domain Specific

In this section we describe the specific approaches taken for our submitted runs for the Domain Specific track. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

Although the Cheshire II system uses the XML structure of documents and extracts selected portions of the record for indexing and retrieval, for the submitted runs this year we used only a single index containing the entire content of the document. This year we did not use the Entry Vocabulary Indexes (search term recommender) that were used by Berkeley in previous years (see [1]), this certainly had an impact on our results, seen in a large drop in average precision when compared to similar runs using these query expansion strategies.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs, however, did *not* use decompounding in the indexing and querying processes to generate simple word forms from compounds (actually we tried, but there was a bug that failed to match any compounds in our runs). This is another aspect of our indexing for this year's Domain Specific task that reduced our results relative to last year.

3.1 Search Processing

Searching the Domain Specific collection used Cheshire II scripts to parse the topics and submit the title and description elements from the topics to the "topic" index containing all terms from the documents. For the monolingual search tasks we used the topics in the appropriate language (English, German, or Russian), and for bilingual tasks the topics were translated from the source language to the target language using SYSTRAN (via Babelfish at Altavista.com) or PROMT via the PROMT web interface. We believe that other translation tools provide a more accurate representation of the topics for some languages (like the L&H P.C. translator used in our GeoCLEF entries) but that was not available to us for our official runs for this track this year. Wherever possible we used both of these MT systems and submitted the resulting translated queries as separate runs. However, some translations are only available on one system or the other (e.g., German *Rightarrow* Russian is only available in PROMT). All of our runs for this track used the TREC2 algorithm as described above with blind feedback using the top 10 terms from the 10 top-ranked documents in the initial retrieval.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the submitted bilingual and monolingual runs for both English and German are shown in Table 1.

Table 1. Submitted Domain Specific Runs

Run Name	Description	translation	MAP
BERK_BI_ENDE_T2FB_B	Biling. English $\Rightarrow$ German	Babelfish	0.23658
BERK_BI_DEEN_T2FB_B	Biling. German $\Rightarrow$ English	Babelfish	0.33013*
BERK_BI_DEEN_T2FB_P	Biling. German $\Rightarrow$ English	PROMT	0.31763
BERK_BI_RUEN_T2FB_B	Biling. Russian $\Rightarrow$ English	Babelfish	0.32282*
BERK_BI_DERU_T2FB_P	Biling. German $\Rightarrow$ Russian	PROMT	0.10826*
BERK_BI_ENRU_T2FB_B	Biling. English $\Rightarrow$ Russian	Babelfish	0.11554
BERK_BI_ENRU_T2FB_P	Biling. English $\Rightarrow$ Russian	PROMT	0.16482**
BERK_MO_DE_T2FB	Monoling. German	none	0.39170
BERK_MO_EN_T2FB	Monoling. English	none	0.41357
BERK_MO_RU_T2FB	Monoling. Russian	none	0.25422**
BERK_MU_DE_T2FB_B.CMBZ	Multiling. from German	PROMT and Babelfish	0.04674
BERK_MU_EN_T2FB_B.CMBZ	Multiling. from English	Babelfish	0.07534**

Table 1 shows all of our submitted runs for the Domain Specific track. Although Berkeley's results for these tasks are fairly high, compared to the overall average MAP for all participants, not much can be claimed for this since there were very few submissions for the Domain Specific track this year, and for many tasks (Multilingual, Monolingual Russian, Bilingual $X \Rightarrow$ Russian, and Bilingual $X \Rightarrow$ English) Berkeley was the *only* group that submitted runs. A few observations concerning translation are worth mentioning. First is that where we have comparable runs using different translations, all of the processing *except* for the topic translation was identical. Thus, it is obvious that for translations from German to English, Babelfish does a better job, and for English to Russian PROMT does a better job. We could not get PROMT (online version) to successfully translate the Russian topics to English apparently due to invalid character codes in the input topics (which were not detected by Babelfish).

It is perhaps more interesting to compare our results with last year's result, where Berkeley was also quite successful (see [1]). This year we see considerable improvement in all tasks when compared to the Berkeley1 submissions for 2005 (which used the same system as this year, but with different algorithms). However, none of this years runs for German Monolingual or Bilingual tasks approached the performance seen for last year's Berkeley2 runs. Because we did no decompounding for German, nor did we do query expansion using EVMs this year, it seems a reasonable assumption that the lack of those is what led to this relative worse performance.

The results of these comparisons with our 2005 Domain Specific results are shown in Table 2. The only exception to the large percentage improvements

Table 2. Comparison of MAP with 2005 Berkeley1 and Berkeley2

Task Description	2005 Berk1	2005 Berk2	2006 MAP	Pct. Diff from Berk1	Pct. Diff from Berk2
Bilingual English⇒German	0.1477	0.4374	0.2366	+60.19	-45.91
Bilingual German⇒English	0.2398	0.4803	0.3301	+37.66	-31.27
Bilingual Russian⇒English	0.2358	–	0.3228	+36.90	
Bilingual German⇒Russian	0.1717	0.2331	0.1083	-36.92	-53.54
Bilingual English⇒Russian	0.1364	0.1810	0.1648	+20.82	-8.95
Monolingual German	0.2314	0.5144	0.3917	+69.27	-23.85
Monolingual English	0.3291	0.4818	0.4136	+25.68	-14.16
Monolingual Russian	0.2409	0.3038	0.2542	+5.52	-16.33
Multilingual from German	0.0294	–	0.0467	+58.84	–
Multilingual from English	0.0346	–	0.0753	+117.63	–

seen by our best 2006 runs over the Berkeley1 2005 runs is found in Bilingual German⇒Russian. We suspect that there were translation problems for the German topics using PROMT, but we lack sufficient language skills in each language to understand what all these problems were. But we did observe that a large number of German terms were not translated, and that many spurious additional characters appeared in the translated texts. Another difference worth noting is that the Berkeley2 group used the L&H PowerTranslator software for its English to X translations, while this year we used only Babelfish and PROMT.

5 Conclusions

Given the small number of submissions for the Domain Specific track this year, we wonder about the viability of the track for the future. Berkeley's runs this year did not build on the successes of the Berkeley2 group from 2005, but instead worked only to establish a new baseline set of results for future for retrieval processing without query expansion using EVMs or thesaurus information. This obviously hurt our comparable overall performance in tasks with submissions from elsewhere. We did, however, see a marked improvement in performance for our specific system (Cheshire II was also used for Berkeley1 last year) with the improvements to our probabilistic retrieval algorithms developed after last year's submissions. We suspect that with the further addition of decompounding for German, and the use of EVMs and thesaurus expansion we can match or exceed the performance of the Berkeley2 runs last year.

References

1. Petras, V., Gey, F., Larson, R.: Domain-specific CLIR of English, German and Russian using fusion and subject metadata for query expansion. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 226–237. Springer, Heidelberg (2006)

Monolingual Retrieval Experiments with a Domain-Specific Document Corpus at the Chemnitz University of Technology

Jens Kürsten and Maximilian Eibl

Chemnitz University of Technology
Faculty of Computer Science, Chair Media Informatics
Straße der Nationen 62
09107 Chemnitz, Germany
jens.kuersten@informatik.tu-chemnitz.de,
eibl@informatik.tu-chemnitz.de

Abstract. This article describes the first participation of the Chair Media Informatics of the Chemnitz University of Technology in the Cross Language Evaluation Forum. An experimental prototype is introduced which implements several methods of optimizing search results. The configuration of the prototype is tested with the CLEF training data. The results of the *Domain-Specific Monolingual German* task suggest that combining the suffix stripping stemming and the decompounding approach is very useful. Also, a local document clustering (LDC) approach used to improve the query expansion (QE) based on pseudo-relevance feedback (PRF) seems to be quite beneficial. Nevertheless, the evaluation of the English task using the same configuration suggests that the qualities of the results are highly speech dependent.

Keywords: Evaluation, Pseudo-Relevance Feedback, Local Clustering, Data Fusion.

1 Introduction

In our first participation in the *Cross-Language Evaluation Forum* we focused on the development of a retrieval system prototype which provides a good baseline for the *Monolingual GIRT tasks*. Furthermore, we tried to improve the baseline results with several optimisation approaches. The outline of this paper is organized as follows. Section 2 introduces the concepts used to achieve good baseline results and Sect. 3 describes the concepts used to optimise those baseline results. Sections 4 and 5 deal with the configurations and results of our officially submitted and of further runs. In Sect. 6 the results are discussed with respect to our expectations.

2 Improving the Baseline Results

We used the *Apache Lucene API* [1] as core to implement well known concepts of information retrieval. Our prototype mainly consists of the following components:

(a) filters for indexing: lowercase filter, stop word filter and stem filter, (b) several algorithms for query construction and modification. The weighting scheme is embedded in the Lucene core and is defined in its documentation.

In our primary experiments with the training data provided by the *CLEF campaign* we concentrated on the *Domain-Specific Monolingual German task*. The components we modified to achieve improved baseline results are introduced in the following subsections.

2.1 Construction of Useful Queries

As we want to participate with *automatic* runs, we have to consider the *CLEF* guideline [2]. It states that all fields of the GIRT corpus can be used for *automatic retrieval*. Thus, we tested several combinations of the given fields and experienced that using the *title*, *abstract*, *controlled-term* and *classification-text* fields is beneficial. Also, the process of constructing the resulting query has to be improved. Well known methods to combine terms in a query formulation procedure are span queries, fuzzy queries and range queries. We experimented with phrase queries without a significant success and with span queries and achieved a slight improvement of precision.

Table 1. Results for the *CLEF* 2005 training data

<i>query construction</i>	<i>map</i>	<i>av. recall</i>
standard	0.3262	0.6458
spans	0.2644	0.3316
standard + spans	0.3621	0.6506
standard + spans + topic analysis	0.3811	0.6674

Finally, the removal of terms within the topics which are not beneficial for retrieval quality could improve the quality of the retrieval results. Therefore, our topic analyzer tries to identify those terms and removes them automatically. As shown in Tab. 1, a significant improvement of both precision and recall was achieved by combining the explained techniques.

2.2 Different Stemmers for German

We compared two different stemming approaches for German. The first one is a part of the *Snowball Project* [3] and the second one was implemented at Chemnitz University of Technology by Stefan Wagner [4]. The development of the second stemmer aimed at the improvement of the effectiveness of our retrieval results. With the decompounding approach every term is divided into a number of syllables. These syllables are treated by the system as tokens. Each token

then can consist of syllables again. The implemented compounding algorithm has two options: (a) remove flexion and (b) avoid overstemming. With the first option we aim at removing flexions from the original term and with the second one we try to remove some letters to avoid the problem of overstemming. Table 2 compares the results of the two stemming approaches and their algorithms based on the *CLEF* 2005 training data: both stemming approaches outperform the non-stemming approach. Furthermore, the compounding approach achieves the best results for the tested topic set.

Table 2. Different stemming algorithms tested with the *CLEF* 2005 training data

<i>stemming algorithm</i>	<i>map</i>	<i>av. recall</i>
none	0.2971	0.5966
plain german compounder	0.4048 (+36.3%)	0.7923 (+32.8%)
snowball german2	0.3811 (+28.3%)	0.6674 (+11.9%)

3 Concepts to Optimise the Monolingual Baseline Results

In this section we introduce and describe the concepts we used to optimise our baseline results. Three different optimisation approaches and their possible combinations were tested. The first subsection takes a look at the query expansion (QE) approach. Thereafter, local clustering (LC) is used to get further improvements. Finally, different algorithms of the data fusion (DF) approach were used to combine the results of the two stemming algorithms from Sect. 2.2.

3.1 Query Expansion

QE is one of the most widely used techniques to improve precision and/or recall of information retrieval systems. The main distinction of the QE approaches is whether manual (user) intervention is needed or not. In our experiments for *automatic retrieval* we used the pseudo-relevance feedback (PRF) approach. Unfortunately, the application of PRF is not useful in all cases. The reason for that is the possible constellation that no or only a few actually relevant documents are used for QE.

Again, we tried to keep things simple and implemented the PRF strategy as follows. We used the top k documents of an initial retrieval run. From those k documents, where $k = 10$ turned out to be a good choice, only terms with an occurrence frequency of at least n were used for QE. Thereby, it can be ensured that only terms from those documents are used that are term correlated and related to the topic anyway. In our experiments we found out that $n = 3$ seems to be a good value for this QE strategy. In further experiments we tried to improve the performance by executing the feedback algorithm repeatedly, but we

observed that only using it twice could be useful. Higher numbers of iterations did not improve retrieval performance; in fact they even deteriorated performance. The results of our successful experiments with QE are summarized in Tab. 3 and compared to a retrieval run without query expansion (first row). Again, we used the *CLEF* 2005 training data for testing.

Table 3. QE based on PRF tested with the *CLEF* 2005 training data

# docs for PRF	min term occ.	iterations	map	av. recall
0	0	0	0.3811	0.6674
10	3	1	0.4556 (+19.5%)	0.8091 (+21.2%)
10	3	2	0.4603 (+20.8%)	0.8154 (+22.2%)

3.2 Local Clustering

This section describes how we applied LC in the QE procedure. The aim of the appliance of clustering was again the improvement of precision and/or recall.

Principally, one has to distinguish between non-hierarchical and hierarchical algorithms for clustering. We tested both concepts to find out what advantages and disadvantages they have in an IR application. The non-hierarchical algorithms are all heuristic in nature because they require a priori decisions, e.g. about cluster size and cluster number. For that reason, these algorithms are only approximating clusters. Nevertheless, the most commonly used non-hierarchical clustering algorithm called k-means was implemented.

Some papers on k-means like algorithms have appeared in the last few years. For example Steinbach et al. compared some of the most widely used approaches with an algorithm they developed [5]. A short overview on non-hierarchical clustering is also given in [6], but the main part of the paper is on hierarchical clustering. In [7] a review of research on hierarchical document clustering is provided.

As already mentioned, we tried to apply clustering in the QE procedure. Therefore, we used the following configurations: (a) *document clustering (LDC)*: using the top k documents returned from clustering the top n documents of an initial retrieval, (b) *term clustering (LTC)*: using the terms returned from clustering the terms of the top n documents of an initial retrieval, (c) *LDC + PRF*: combining (a) with PRF from section 3.1 and (d) *LTC + PRF*: combining (b) with PRF. For LDC our experiments showed, that clustering the top $n = 50$ documents of an initial retrieval provides the best results.

The LTC approach is quite different from LDC, because the objects to be clustered are not the top k documents themselves, but the terms of those top k documents. To get more reliable term correlations, we used the top $k = 20$ documents for LTC. A brief description of some methods for LTC is given in [8].

Table 4 lists the results of our four clustering configurations by testing each of them with the *CLEF* 2005 training data. Clustering for QE is compared to

our baseline results (first row) and the PRF approach (second row). The results show that LDC can achieve a small improvement at precision and recall, when it is combined with the PRF approach described in Sect. 3.1. Furthermore, it is obvious that LTC did not improve precision or recall.

Table 4. Different LC techniques for QE tested with the *CLEF* 2005 training data

	<i>clustered objects map</i>	<i>av. recall</i>
-	0.3811	0.6674
-	0.4603 (+20.8%)	0.8154 (+22.2%)
terms (b)	0.4408 (+15.7%)	0.8024 (+20.2%)
terms (d)	0.4394 (+15.3%)	0.7968 (+19.4%)
docs (a)	0.4603 (+20.8%)	0.7949 (+19.1%)
docs (c)	0.4726 (+24.0%)	0.8225 (+23.2%)

3.3 Data Fusion

This section gives a short overview of the result list fusion approaches we implemented to combine the different indexing schemes described in Sect. 2.2. Data fusion has been in the interest of many researchers in the IR field. In [9] different operators for result list fusion problem have been compared. In this report *summarizing the RSV values* performed best. Savoy [10] also compared some data fusion operators and he additionally suggested the *z-score* operator, which performs best in his study. Another fusion operator called *norm-by-top-k* was introduced by Lin and Chen in their report on merging mechanisms for multi-lingual information retrieval [11].

Beside the widely known *round robin* and *raw score* fusion operators, the fusion operators depicted in Tab. 5 were implemented.

Table 5. Different data fusion operators tested with the *CLEF* 2005 training data

<i>fusion operator map</i>	<i>av. recall</i>
-	0.4048 0.7923
round robin	0.4296 0.7875
raw score	0.4070 0.7938
sum RSV	0.4267 0.7938
norm RSV	0.4311 0.8057
norm max	0.4256 0.7867
norm by top-k	0.4297 0.7878
z-score	0.4324 0.8050

All the tested fusion operators improved the precision of our primary results as can be seen in Tab. 5. Furthermore, one can observe that *z-score* performed

best and *norm-RSV* performed almost as good as *z-score*. Another interesting observation is that the simple *round robin* approach also achieved very good results. The only fusion operator that performed poor compared to the others was *raw score*. The reason for that is the missing normalization of the used result lists.

4 Configurations and Results of Official Runs

This section provides an overview of the configuration of the runs we submitted for the *Domain-Specific Monolingual tasks* in german and english. All of our experiments used the weighting scheme embedded in the Lucene core, which is a *tf*idf* weighting scheme (see [1]).

Table 6 summarizes the configurations of our official runs. For the QE two different PRF approaches were used: (a) PRF1, single step PRF and (b) PRF2, two-step iterative PRF.

Table 6. Configurations and results of official runs

<i>run identifier</i>	<i>lang</i>	<i>stemmer</i>	<i>QE</i>	<i>data fusion map</i>	
TUCMIgirtde1	de	snowball(german2)	PRF2	none	0.5122
TUCMIgirtde2	de	snowball(german2)	PRF2 + LDC	none	0.5202
TUCMIgirtde3	de	snowball(german2)	PRF2 + LTC	none	0.5136
TUCMIgirtde4	de	snowball(german2), decompounder	PRF2	sum RSV	0.5454
TUCMIgirten1	en	snowball(porter)	PRF2	none	0.3510
TUCMIgirten2	en	snowball(porter)	PRF2 + LDC	none	0.3553
TUCMIgirten3	en	snowball(porter)	PRF2 + LTC	none	0.3450
TUCMIgirten4	en	snowball(porter)	PRF1	none	0.3538

The results show, that PRF in combination with LDC achieves slighth improvements of the precision, whereas the combination of PRF and LTC is less promising, because it deteriorates precision for the english task. The data fusion of the two different stemming approaches achieves the best results.

5 Configurations and Results of Additional Runs

In this section we summarize further experiments, which were not officially submitted. The configurations and results for those experiments are given in Tab. 7. Again, we used the concepts PRF and LDC for QE, as well as their combinations for the data fusion.

The results show that the revised version of the algorithm for LDC performs better than the former version. Also, the data fusion of the runs *TUCMIgirtde5* and *TUCMIgirtde7* with the use of the *z-score* operator leads to a further improvement of the precision.

Table 7. Configurations and results of additional runs

<i>run identifier</i>	<i>lang stemmer</i>	<i>QE</i>	<i>data fusion map</i>	
TUCMIgirtde5 de	snowball(german2)	PRF2 + LDC	none	0.5278
TUCMIgirtde6 de	decompounder	PRF2	none	0.4670
TUCMIgirtde7 de	decompounder	PRF2 + LDC	none	0.4862
TUCMIgirtde8 de	snowball(german2), decompounder	PRF2 + LDC	z-score	0.5632

6 Conclusions

Generally speaking, the results of our experiments were as expected. The result list fusion approach for the *Domain-Specific Monolingual German* task clearly outperformed our other experiments. Moreover, LDC in combination with the PRF approach achieves slight better results than PRF alone or PRF in combination with LTC.

The results of the *Domain-Specific Monolingual English* task also show that PRF in combination with LDC performs best. However, PRF with LTC worked worst and in comparison to the german task the differences in the results for our configurations were very small.

Concluding our experiments one can say, that for german text retrieval, combining the suffix stripping stemming and the decompounding approach is very useful and further experiments with other languages should be done to determine if one can achieve similar results with them. Also, the LDC approach for the improvement of QE seems to be quite beneficial, but again further experiments, especially with different corpora, should be carried out to detect, if the improvement works only within this (domain-specific) environment or not.

References

1. The Apache Software Foundation: Lucene. Retrieved August 10, 2006 from the World Wide Web (1998-2006), <http://lucene.apache.org>
2. CLEF: Guidelines for Participation in CLEF 2006 Ad-Hoc and Domain-Specific Tracks. Retrieved August 10, 2006 from the World Wide Web (restricted access, 2006), <http://www.clef-campaign.org/delos/clef/protect/guidelines06.htm>
3. Porter, M.: The Snowball Project. Retrieved August 10, 2006 from the World Wide Web (2001), <http://snowball.tartarus.org>
4. Wagner, S.: A German Decompounder Retrieved August 10, 2006 from the World Wide Web (2005), <http://www-user.tu-chemnitz.de/~wags/cv/clr.pdf>
5. Steinbach, M., Karypis, G., Kumar, V.: A Comparison of Document Clustering Techniques, University of Minnesota, Technical Report # 00034 (2000)
6. Rasmussen, E.: Clustering Algorithms. In: Frakes, W.B., Baeza-Yates, R. (eds.) *Information Retrieval -Data Structures and Algorithms*, Prentice Hall, Englewood Cliffs New Jersey (1992)
7. Willett, P.: Recent Trends in Hierarchic Document Clustering. *Information Processing & Management* 24(5), 577–597 (1988)

8. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Pearson Addison-Wesley, Harlow Munich (2005)
9. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proceedings of the 2nd Text Retrieval Conference (TREC2), NIST Special Publication, pp. 215–500 (1994)
10. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval. In: Working Notes for the CLEF 2004 Workshop (2004)
11. Lin, W.-C., Chen, H.-H.: Merging Mechanisms in Multilingual Information Retrieval. In: Working Notes for the CLEF 2002 Workshop (2002)

iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository

Jussi Karlgren¹, Julio Gonzalo², and Paul Clough³

¹ SICS, Kista, Sweden

² UNED, Madrid, Spain

³ University of Sheffield, Sheffield, UK

Abstract. This paper summarizes the task design for iCLEF 2006 (the CLEF interactive track). Compared to previous years, we proposed a radically new task: searching images in a naturally multilingual database, Flickr, which has millions of photographs shared by people all over the planet, tagged and described in a wide variety of languages. Participants were expected to build a multilingual search front-end to Flickr (using Flickr's search API) and study the behaviour of the users for a given set of searching tasks. The emphasis was put on studying the process, rather than evaluating its outcome.

1 Introduction

iCLEF (the CLEF interactive track) has been devoted, since 2001, to study Cross-Language Retrieval from a user-centered perspective. The aim has always been to investigate real-life cross-language searching problems in a realistic scenario, and to obtain indications on how best to aid users in solving them. iCLEF experiments have investigated the problems of foreign-language text retrieval, question answering and image retrieval, including aspects such as query formulation, translation and refinement, and document selection. The focus has always been on improving the outcome of the process in terms of a classic notion of relevance, and the target collection has always consisted of news texts in languages foreign to the user. Finally, the task has always involved the comparison of a reference system with a contrastive system, combining users, topics and systems with a latin-square design to detect system effects and filter out other effects.

Although iCLEF in only a few years of activity has established the largest collected body of knowledge on the topic of interactive Cross-Language Retrieval, the experimental setup has proven limited in certain respects:

- The search task itself was unrealistic. News collections are comparable across languages, and most of the pertinent information tends to be available in the user's native language; why would a user search for this information in an unknown language? Translating this problem to the WWW, it would be like asking a Spanish speaker to find information about the singer Norah Jones

in English, in spite of the fact that there are over 150,000 pages in Spanish about her¹.

- The target notion of “relevance” does not cover all aspects that make an interactive search session successful.
- The latin-square design imposed heavy constraints on the experiments, making them costly and with a limited validity (the number of users was necessarily limited, and statistically significant differences were hard to obtain).

In order to overcome these limitations, we have decided to propose a new, pilot framework for iCLEF which has two essential features:

- We have chosen Flickr (the popular photo sharing service) as the target collection. Flickr is a large-scale, web-based image database serving a large social network of WWW users. It has the potential to offer both challenging and realistic multilingual search tasks for interactive experiments.
- We want to use the iCLEF track to explore alternative evaluation methodologies for interactive information access. For this reason, we have decided to fix the search tasks, but to keep the evaluation methodology open. This allows each participant to contribute with their own ideas about how to study interactive issues in cross-lingual information access.

2 The Multilingual Reality of the Internet

Estimates of online users’ population per language have a wide range and a considerable margin of error, but by comparing estimates of user population with assessed presence of material by language, it can fairly be said that materials in many languages, notably English and other North European ones, are currently over-represented by user count and many other languages are under-represented — compared to the numbers of first-language users. Both user population statistics and available materials change apace and all numerical estimates age rapidly: whichever way the statistics are collected, it is certain that the proportion of primarily English users is shrinking (from 36% in 2004 to 31% in 2006) and that other languages are growing in usage with the entrance of large user populations that have little or no command of English. These trends are strong motivations for cross-lingual retrieval, which of course will be found to be most useful for non-linguistic materials such as images.

3 The Target Collection: Flickr

The majority of Web image search is text-based, and the success of such approaches often depends on reliably identifying relevant text associated with a particular image. Flickr is an online tool for managing and sharing personal photographs, and currently contains over thirty million freely accessible images². These are updated daily by a large number of registered users and available to all web users.

¹ Google.com results as of 11 August 2006.

² As of August, 2006.

3.1 Photographs in the Collection

Flickr provides both private and public image storage, and photos which are shared (around 5 million) can be protected under a Creative Commons (CC) licensing agreement (an alternative to full copyright). Images from a wide variety of topics can be accessed through Flickr, including people, places, landscapes, objects, animals, events, etc. This makes the collection a rich resource for image retrieval research.

There were two possibilities to use the collection: reaching an agreement with Flickr to get a subset of their database, or simply using Flickr's public API to interact with the full database. The first option is more attractive from the point of view of system design, because it is possible to obtain collection statistics to enhance the search (for instance, tf-idf weights, term suggestions and term translations adapted to the collection) and because it gives total control on the search mechanism. A crucial advantage of having the collection locally stored is enabling the possibility of doing content-based retrieval.

The second option, by contrast, by design reflects the dynamic nature of open databases and allows users to interact with a larger and more current target database - which is more realistic and thus preferable. While recall-oriented success measures are ill defined, collection frequencies more complex to calculate, and reproducibility of results can be called into question, this option relieves organisers and participants alike from cumbersome administration and distribution of test collections. In any case, until an agreement with Flickr is reached, using Flickr's API to access the full Flickr database was the only choice available for this first pilot task.

As of October 2005, 1.2 million Flickr users added around 200,000 images a day to the collection³. In order to keep the collection somewhat more constant across the experiments, we decided to restrict the experiments to images uploaded before 21 June 2006 (immediately before iCLEF experiments began).

3.2 Annotations

In Flickr, photos are annotated by authors with freely chosen keywords ("tags") in a naturally multilingual manner: most authors use keywords in their native language; some combine more than one language. User tags may describe anything related to the picture, including themes, places, colours, textures and even technical aspects on how the photograph was taken (camera, lens, etc.). Some tags become naturally "standardized" among subsets of Flickr users, in a typical process of so-called "folksonomies" [5].

In addition, photographs have titles, descriptions, collaborative annotations, and comments in many languages. Figure 1 provides an example photo with multilingual annotations.

Photos can also be submitted to online discussion groups. This provides additional metadata to the image which can also be used for retrieval. An explore

³ <http://www.wired.com/news/ebiz/0,1272,68654,00.html>

outono/inverno

ALL
SIZES



começando a série "velha, mas enxuta" uma da BR 116 num dia assim como o de hoje frio e chuvoso.

vontade de fazer esse mesmo percurso com mais calma dia desses, hein Hartmann?

grande abraço e bom feriado,

[750 x 1125](#)

de volta ao mundo dos Interestingness!

Uploaded on June 15, 2006
by [corbata1982](#)

[corbata1982's
photostream](#)

This photo also belongs to:

+ [menu \(Set\)](#)

+ [tus faves \(Set\)](#)

+ [mis faves \(Set\)](#)

+ [* La tengo y te reto *
\(Pool\)](#)

+ [* TE DESAFIO * \(Pool\)](#)

+ [1 million photos \(Pool\)](#)

Tags

BR116
estrada
way
via
caminho
árvores
trees
tree
árbol
arboles
árvore
chuva
rain
lluvia
frio

Fig. 1. An example Flickr image, with title, description, classified in three sets (user-defined) and three pools (community shared), and annotated with more than 15 English, Spanish and Portuguese tags

utility provided by Flickr makes use of this user-generated data (plus other information such as viewing statistics and ratings) to define an “interestingness”⁴ view of images.

3.3 Flickr’s Search API

While Flickr has many modes in which users can explore the photo collection, its search capabilities are rather simplistic. One can choose between searching tags

⁴ <http://www.flickr.com/explore/interesting>.

only, or full text search (title, description and tags). In the tag searching mode there are two options: conjunctive (all keywords) or disjunctive (any keyword). In the full text search, only the conjunctive mode is provided⁵. This is a serious restriction when performing cross-language searches that had to be taken into account by participants in the task.

During the experiments, the API has offered fast and stable response times.

4 The Task

Images are naturally language independent and often successfully retrieved with associated texts. This has been explored as part of ImageCLEF (Clough et al, 2005) for areas such as information access to medical images and historic photographs. This makes images specially attractive as a scenario where cross-language search arises more naturally.

The way in which users search for images provides an interesting application for user-centered design and evaluation. As an iCLEF task, searching for images from Flickr presents a new multilingual challenge which, to date, has not been explored. Challenges include:

- Different types of associated text, e.g. tags, titles, comments and description fields.
- Collective classification and annotation using freely selected keywords (folksonomies) resulting in non-uniform and subjective categorization of images.
- Fully multilingual image annotation, with all widely-spoken languages represented and mixed-up in the collection.
- Large number of images available on a wide variety of topics from different domains.

Given the multilingual nature of the Flickr annotations, translating the user's search request would provide the opportunity of increasing the number of images found and make more of the collection accessible to a wider range of users regardless of their language skills. The aim of iCLEF using Flickr will be to determine how cross-language technologies could enhance access, and explore the user interaction resulting from this.

The experiment consists of three search tasks, where users may employ a maximum of twenty minutes per task. We have chosen three tasks of a different nature:

- Topical ad-hoc retrieval: *Find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside.*
- Creative open-ended retrieval: *Find five illustrations to the article "The story of saffron"* (see the text in Figure 2).

⁵ At the time of writing this overview, Flickr has started providing full boolean search, which can be exploited in future experiments.

- Visually oriented task: *What is the name of the beach where this crab is resting?* (along with a picture of a crab lying in the sand, see Figure 3). The name of the beach is included in the Flickr description of the photograph, so the task is basically finding the photograph, which is annotated in German – a fact the users is unaware of.

Abruzzo Dishes: The Story of Saffron

Obtained from the dried and powdered stem of the ‘*croccus sativus*’ which grows on the Navelli Plain in the Province of L’Aquila, saffron is considered by many to be the single most representative symbol of the traditional products of Abruzzo.

An essential ingredient in Risotto Milanese, the spice also crops up in many other dishes across Italy. For example, the fish soup found in Marche, south of the Monte Conero, contains saffron for its red coloring in place of the more traditional tomato. This coloring property is also widely appreciated in the production of cakes and liqueurs and for centuries by painters in the preparation of dyes. Its additional curative powers have long been known to help digestion, rheumatism and colds.

How a flower of Middle Eastern origin found a home in this unfashionable corner of Italy can be attributed to a priest by the name of Santucci who introduced it to his native home 450 years ago. Following his return from Spain at the height of the Inquisition, his familiarity with Arab-Andalusian tradition convinced him that the cultivation of the plant was possible in the plains of Abruzzo, and so it proved.

Nevertheless, even today the harvesting of saffron is hard and fastidious work with great skill needed to handle the stems without damaging the product inside or allowing contamination from other parts of the plant. The area of cultivation in the region is strictly limited to 8 hectares of land. A sad reduction from the 430 hectares cultivated at the turn of the last century.

Together with the labor intensiveness of the production and the care and patience involved in gathering and drying the flower, the cost of Abruzzese saffron is high relative to its competitors from the Middle East and Iran. Yet all are agreed it possesses superior aromatic qualities and remains the preferred choice in gourmet cooking. It is so good that the saffron from the area is practically all exported.

Anyone interested in buying the end product should note that although sachets of saffron powder can be purchased, the real thing should only be bought as the characteristic dried fine stems.

The saffron is grown in an area comprising the comune of Navelli, Civitaretenga, Camporciano, San Pio delle Camere and Prata D’Ansionia.

Fig. 2. Creative task: Find five illustrations to this text

All tasks can benefit from a multilingual search: Flickr has photographs of European parliament buildings described in many languages, photographs about the Abruzzo area and saffron are only annotated in certain languages, and the crab photograph can only be found with German terms. At the same time, the nature of each task is different from the others. The “European parliaments” topic is biased towards recall, and one can expect users to stop searching only when the twenty minutes expire. The text illustration task only demands five



Fig. 3. Visually oriented task: *What is the name of the beach where this crab is resting?*

photographs, but it is quite open-ended and very much depending on the taste and subjective approach of each user; we expect the search strategies to be more diverse here. Finally, the “find the crab” task is more of a known-item retrieval task, where the image is presumed to be annotated in a foreign language, but the user does not know which one; the need for cross-language search and visual description is more acute here.

Given that part of the experience consisted of proposing new ways of evaluating interactive Cross-Language search, we did not prescribe any fixed procedure or measure for the task. To lower the cost of participation, we provided an Ajax-based basic multilingual interface to Flickr, which every participant could use as a basis to build their systems.

5 Experiments

Fourteen research groups officially signed in for the task, more than in any previous iCLEF edition. However, only the three organizing teams (SICS, U. Sheffield and UNED) submitted results (the worst success rate in iCLEF campaigns). Finding the reasons is left to discussion along the workshop; perhaps the fact that the task is only half-cooked made people feel insecure about what or how to measure. We hope that the results obtained by the three organizing groups will encourage a much broader participation next year.

This is a brief summary of the experiments performed at iCLEF 2006:

- UNED [1] measured the attitude of users towards cross-language searching when the search system provides the possibility (as an option) of searching cross-language, using a system which allowed for three search modes: “no translation”, “automatic translation” (the users chooses the source language and the target languages, and the system chooses a translation for every word in the query) and “assisted translation” (like the previous, but now the user can change the translation choices made by the system). Their results over 22 users indicate that users tend to avoid translating their query into unknown languages, even when the results are images that can be judged visually.
- U. Sheffield (and IBM) [2] experimented with providing an Arabic interface to Flickr in an attempt to increase the amount of material accessible to the Arabic online community. An Arabic-English dictionary was used as an initial query translation step, followed by the use of Babelfish to translate between English and French, German, Dutch, Italian and Spanish. Users were able to modify the English translation if they had the necessary language skills. With the user group testing the system (bilingual Arabic-English students), it was found that these users: preferred to query in English (although liked having the option available to them of formulating the initial queries in Arabic), found the system very helpful and easy to use, and overall were able to search effectively in all tasks provided. Users found viewing photos with results in multiple languages more helpful and important than the initial query translation step from Arabic to English. Users needed to view the image annotations for most tasks, but had no problems doing this for the languages available to them. If non-European languages had been included (e.g. Chinese or Japanese) then users would not have been able to use the annotations effectively. An analysis of the tasks was also carried out and included in the results. Overall it was found that these tasks were not well-suited to Arabic users.
- SICS [3] The experiments carried out at SICS this year centered on user satisfaction and user confidence as target measures for information access evaluation. The users were given the tasks, and after some time were given a terminological display of terms they had made use of together with related terms. This enabled them to broaden their queries: success was not measured in retrieval results but in change of self-reported satisfaction and confidence as related to the pick-up of displayed terms by the user[4].

6 Conclusions

Flickr would certainly seem to provide cross-language research with a highly plausible and well-motivated cross-language search scenario, especially given that results are visual in nature. The goal of iCLEF 2006 has been to investigate this particular cross-language scenario with users on a range of search tasks designed to favour multilingual search. While the number of participants has always been

low in any organised evaluation events involving users, this year number of submissions cannot be viewed as anything but disappointing: the lowest turnout so far in iCLEF. However, the interest as measured by pre-registration was considerable. This year, the lessons learnt are mainly methodological — can target measures based on other notions than “relevance” give interesting results? The submissions to iCLEF do, however, offer some insights into cross-language searching in Flickr and these will be discussed at the CLEF workshop, together with a discussion on how participation better could be encouraged for future evaluation cycles.

Acknowledgments

This work has been partially supported by the Spanish Government under project R2D2/Symbra (TIC2003-07158-C04-02) and by the European Commission, project MultiMatch (FP6-2005-IST-5-033104) and the DELOS Network of Excellence (FP6-G038-507618).

References

1. Artiles, J., Gonzalo, J., López-Ostenero, F., Peinado, V.: Are users willing to search cross-language? an experiment with the flickr image sharing repository. [In This volume] (2006)
2. Clough, P., Al-Maskari, A., Darwish, K.: Providing multilingual access to flickr for arabic users. [In This volume] (2006)
3. Karlgren, J., Olsson, F.: Trusting the results in crosslingual keyword-based image retrieval. [In This volume] (2006)
4. Karlgren, J., Sahlgren, M.: Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering* 11(3), 327–341 (2005)
5. Mathes, A.: Folksonomies-cooperative classification and communication through shared metadata. Technical Report LI590CMC, Computer Mediated Communication, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (2004)

Are Users Willing to Search Cross-Language? An Experiment with the Flickr Image Sharing Repository

Javier Artiles, Julio Gonzalo, Fernando López-Ostenero, and Víctor Peinado*

NLP Group, ETSI Informática, UNED
javart@bec.uned.es, {julio, flopez, victor}@lsi.uned.es

Abstract. This paper summarizes the participation of UNED in the CLEF 2006 interactive task. Our goal was to measure the attitude of users towards cross-language searching when the search system provides the possibility (as an option) of searching cross-language, and when the search tasks can clearly benefit from searching in multiple languages.

Our results indicate that, even in the most favorable setting (the results are images that can be often interpreted as relevant without reading their descriptions, and the system can make translations in a transparent way to the user), users often avoid translating their query into unknown languages.

1 Introduction

CLEF,¹ NTCIR² and TREC³ evaluation campaigns have contributed, along the years, to create an extensive corpus of knowledge on Cross-Language Information Retrieval from an algorithmic perspective. Little is yet known, however, on how users will benefit from Cross-Language retrieval facilities.

iCLEF⁴ (the CLEF interactive track) has been devoted, since 2001, to study Cross-Language Retrieval from a user-centered perspective. Many things have been learned in the iCLEF framework about how a system can best assist users when searching cross-language. But all iCLEF experiments so far were slightly artificial, because users were forced to search in a foreign language. Previous UNED experiments [2], [3], [4], for instance, consisted of native Spanish speakers which were asked to search a news collection written entirely in English. The task was artificial because the contents they were searching were also available in other news collections in Spanish, their native language.

iCLEF 2006 proposes a radically new task, which consists of searching images in a naturally multilingual database, Flickr,⁵ which has millions of photographs

* Authors are listed in alphabetical order.

¹ See <http://www.clef-campaign.org/>.

² See <http://research.nii.ac.jp/ntcir/>.

³ See <http://trec.nist.gov/>.

⁴ See <http://nlp.uned.es/iCLEF>.

⁵ See <http://www.flickr.com>.

shared by people all over the planet, tagged and described in a mixture of most languages spoken on earth.

We have used the iCLEF 2006 task design to find out how users react to a system that provides cross-language search facilities on a naturally multilingual image collection. If searching cross-language is an option of the system (rather than a requisite of an experiment), will users take advantage of this possibility? How will their language skills influence their use of the system? How will the nature of the search task influence the degree of multilinguality they achieve when searching? Will there be an inertial effect from the dominant search mode (exact match, all words conjunctively, no expansion/translation) used by all major search engines? Do they perceive the system as useful for their own search needs?

Rather than focusing on the outcome of the search process, we have therefore focused on the search behavior of users. We have designed a multilingual search interface (a front-end for the Flickr database) where users can search in three modes: *no translation*, *automatic translation* in the languages selected by the user, and *assisted translation*, where users can change the translations initially picked up by the system. Our users have conducted the three search tasks prescribed by the iCLEF design [1], and we have studied (through observations and log analysis) their search behavior and their usage of cross-language search facilities. Finally, we have also contrasted this information with the subjective opinions of the users, stated in a post-experiment questionnaire.

This paper is structured as follows. First, in Section 2, we present the experimental design of the experiment: the test collection used, the tasks proposed and the Flickr front-end implemented by our group. In Section 3, we discuss the results obtained, both comparing search modes (3.1) and target languages used (3.2) and analyzing the observational study (3.3) and the data compiled from the user questionnaires (3.4). Finally, in Section 4, we draw some conclusions.

2 Experiment Design

Our experiment follows iCLEF guidelines [1]. In summary:

Test collection. The collection to be searched is all public photos in Flickr uploaded before 21 June 2006. This date is fixed so that everyone is searching exactly the same collection. This is a collection of more than 30 million photographs annotated with title, description and tags. Tags are keywords freely chosen by users; community usage of tags creates a so-called “folksonomy”.

Access to the collection. The collection can be accessed via Flickr’s search API,⁶ which allows only two search modes: search tags (either all tags in the query or any tag in the query) and full search (search all query terms in title, description and tags). No exact statistics on the collection (images, size, vocabulary, term frequencies) were available for the experiment.

⁶ For further details about the Flickr’s API and its documentation see <http://www.flickr.com/services/api/>.

Search tasks. iCLEF guidelines prescribed at least three tasks to be performed by users, which could employ a maximum of twenty minutes per task:

- Ad-hoc task: *Find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside.*
- Creative task: *Find five illustrations to the article “The story of saffron”, a one-page text about cultivation of saffron in Abruzzo, Italy.*⁷
- Visually oriented task: *What is the name of the beach where this crab is resting?*, along with a picture of a crab lying in the sand.⁸

The name of the beach is included in the Flickr description of the photograph, so the task is basically finding the photograph, which is annotated in German (a fact that the users ignore) and identifying the name of the beach.

All tasks can benefit from a multilingual search: Flickr has photographs of European parliament buildings described in many languages, photographs about the Abruzzo area and saffron are only annotated in certain languages, and the crab photograph can only be found with German terms.

With these constraints, we have designed an experiment that involves:

- 22 users, all of them native Spanish speakers and with a range of skills in other languages.
- A search front-end to Flickr.
- A pre-search questionnaire, asking users about their experience searching images and using Flickr, and about their language skills in the six languages proposed by our interface.
- A post-search questionnaire, asking users about the perceived usefulness of cross-language search facilities and the degree of satisfaction with the results.

The key of the experiment is the design of the search front-end with Flickr. It has three search modes: *no translation*, *automatic translation*, and *assisted translation*. Users can switch between these search modes at will. In both translation modes, users can select Spanish, English, French, Italian, Dutch and German as source language, and any combination of them as target languages.

In the *no translation* mode, the interface simply launches the queries against the Flickr database using its API. Users may choose between searching only the tags, or searching titles, tags and descriptions (full text). In the tags mode, they can select a conjunctive or a disjunctive search mode. In the full text search mode, the search is always conjunctive. These are the original search options with Flickr’s API.

In the *automatic translation* mode, the user can choose the source language and the target languages. The system shows a result box per language, together with the query as it is formed in every language. The user cannot manipulate the translation process directly. As the full text search is always conjunctive in

⁷ The English version of the text is available at <http://nlp.uned.es/iCLEF/saffron.txt>.

⁸ The picture is available at <http://nlp.uned.es/iCLEF/topic3.jpg>.

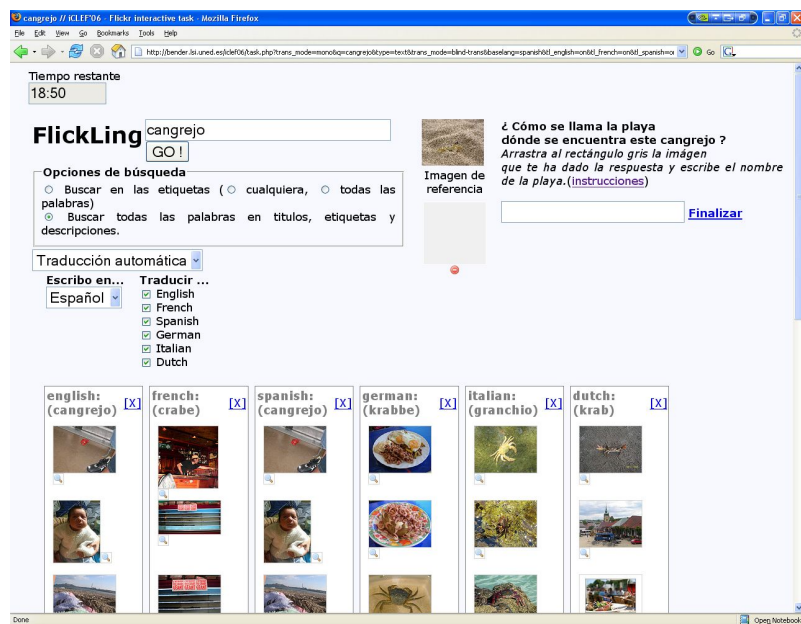


Fig. 1. Dedicated search interface for the “find the crab” task

the Flickr API, we chose to translate each original term in the query with only one term in every target language, to avoid over-restrictive queries.

In the *assisted translation*, the user has the additional possibility of changing the default translations chosen by the system. A code color is used to signal alternative translations coming from the same source term. The user can change the preferred translation for every term as desired; the query is automatically re-launched (only in the language where it was refined) with the new term. As a help to select the most appropriate translation, the system displayed inverse translation when moving the mouse over a translated term.

For the experiment, we have implemented three versions of this search interface which facilitate performing each task and providing results for the user. For instance, Figure 1 shows a snapshot of the dedicated interface for the “find the crab” visually-oriented task. In the left-hand side of the interface, the original crab image is always shown, together with the task description. Immediately below there is an empty box where the user must drag-and-drop the image that provides the answer (name of the beach), a text box where s/he can write the answer, and a “Done” button to end the task. In the upper left corner, a clock indicates how much time is left to complete the task.

For translation we have only used freely available resources. We have built databases for all language pairs from bilingual word lists (with approx. 10,000 entries per language) provided by the Universal Dictionary Project⁹. All entries

⁹ Dictionaries can be downloaded from <http://dicts.info/uddl.php>.

were stemmed using a specific Snowball stemmer¹⁰ for each language. The *automatic translation* mode simply picks up the first translation in the dictionary for every word in the query. If a given word is not found in the database, it remains untranslated.

3 Results and Discussion

3.1 Comparison Among Monolingual, Automatic Translation and Assisted Translation Search Modes

As explained above, our interface implements three different search modes: *no translation*, which launches the query as it is, *automatic translation* where users can select target languages and the systems picks up the first translation for each query term, and *assisted translations*, where users can also select target translations.

Figure 2 shows the proportion of total subjects using every search mode for each task. It seems that the *assisted translation* was the most popular search mode across tasks. It allowed both to select the right translation for each query term when the *automatic translation* was not accurate enough and to learn or remember new equivalent translations.

It is noticeable that most users begin the activities with the *no translation* mode (the system default) but it's discarded relatively soon (mostly, after the first minute) as soon as subjects realize that a translation facility is needed to complete a task. This fact is confirmed in Figure 2a. Specifically, for the “find this crab” task, we can also see how the figures tend to lower as the time goes by: the number of subjects still on search decreased because some users finished the task before the 20-minute limit.

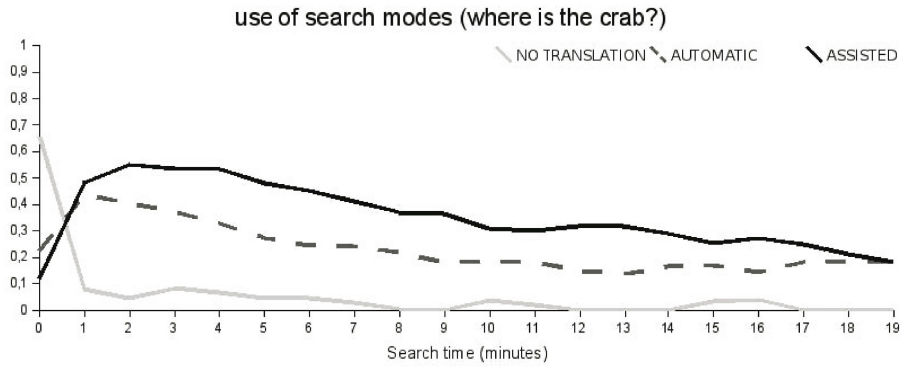
For the “European parliaments” task (Figure 2b), we can see that the number of subjects on search remains stable: most of the users employed the available 20 minutes in order to grab as many photographs as possible.

In the “saffron” task, in Figure 2c, it is interesting to note that the *no translation* mode is higher than in other tasks. This seems to point out that, for a creative activity, some users do not consider essential to retrieve pictures annotated in other languages.

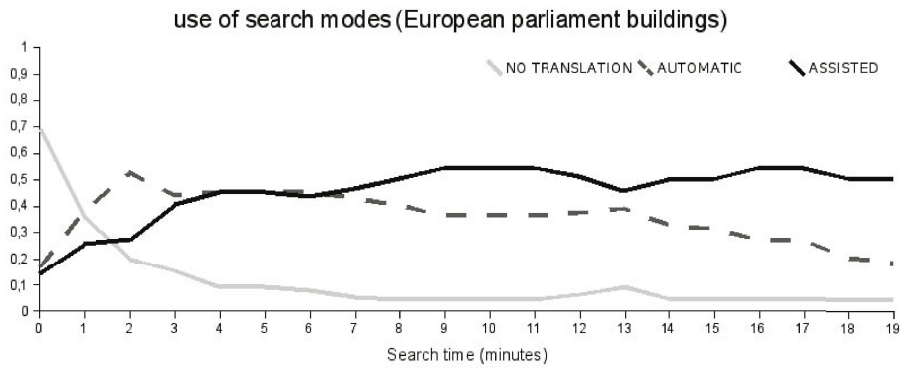
3.2 Comparison of Searches in Native, Familiar and Unknown Languages

In Figure 3 we show the ratio of subjects using a *native*, *passive* or *unknown* language as target during each task. For comparison purposes, we have classified subjects according to the information compiled in the pre-session questionnaires as follows: the users who declared to be native or fluent speakers are represented as “native”. The ones who declared to be able to read but not to write fluently are represented as “passive” speakers. Finally, the ones who claimed not to understand a single word are shown as “unknown”.

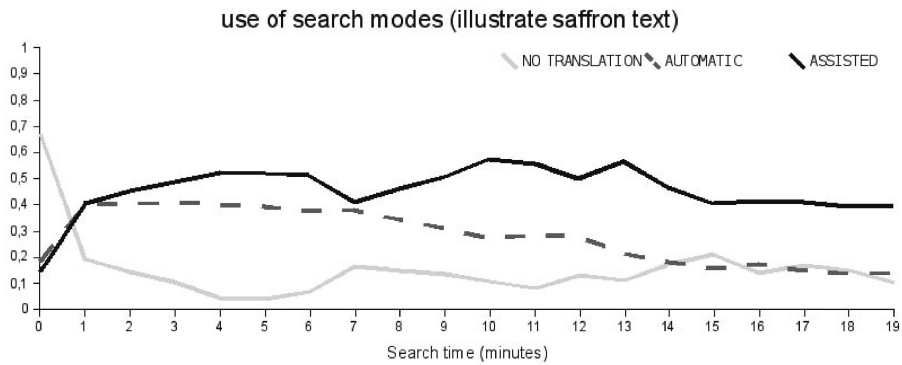
¹⁰ See <http://www.snowball.tartarus.org/> for further information about Snowball.



(a)

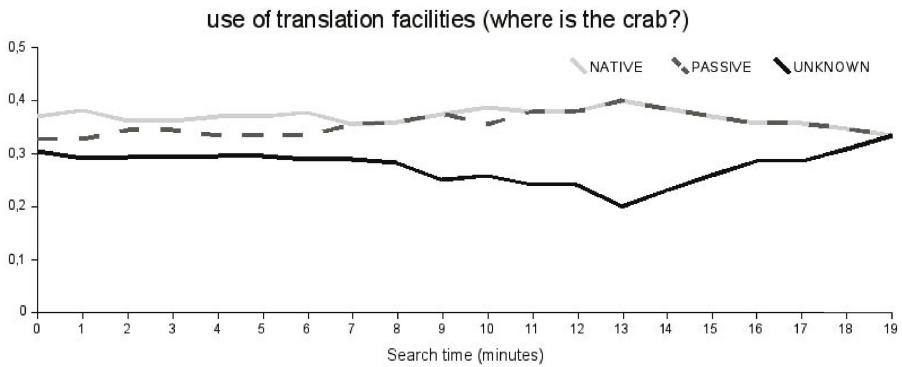


(b)

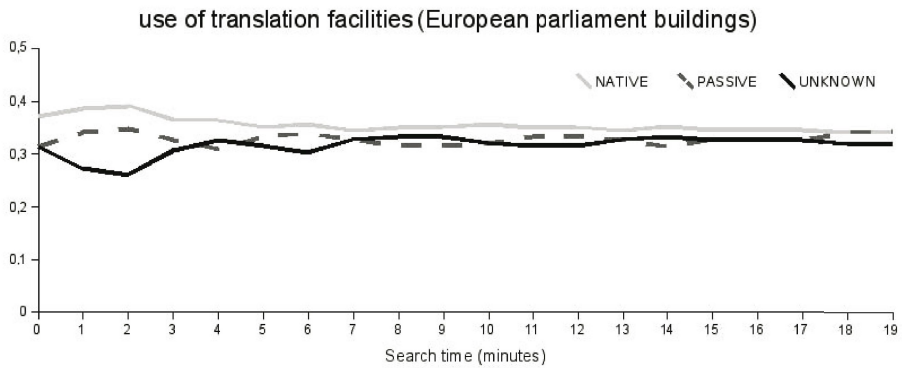


(c)

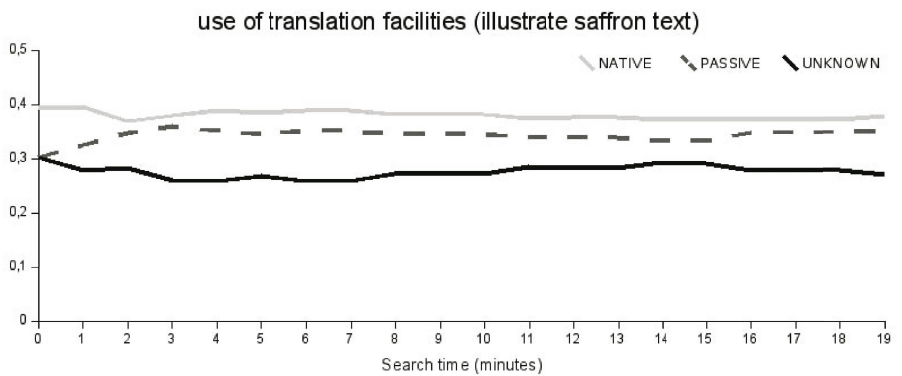
Fig. 2. Use of the search modes implemented in the system: *no translation*, *automatic translation* (users select target languages) and *assisted translations* (users can also select target translations)



(a)



(b)



(c)

Fig. 3. Use of target languages classified according to self-declared user language skills

For the “find this crab” task, as shown in Figure 3a, we can see that users preferred to perform their searches using languages in which they may feel confident rather than using unknown languages. Notice that there is only one picture containing the right answer and, therefore, this is clearly a precision-oriented task. The photograph was annotated in German, which is an unknown language for most of our users. Since they were not willing to search in German, that made the task hard to complete. Indeed, only 9 out of 22 subjects were able to find the crab. As subjects approach the time limit, the use of unknown languages rises up. This can be interpreted as a final attempt to come up with the crab.

On the contrary, the “European parliaments” task can be considered as a recall-oriented activity. Figure 3b shows the same proportion of subjects performing searches over known and unknown target languages, because they tried to grab pictures from any possible source.

Lastly, the “saffron” task is the most creative one. It is hard to evaluate whether our subjects performed well or not. They had to choose photographs related to the content in order to enhance the text. As said in section 3.1, most of the users considered enough to search in known languages (see Figure 3c).

3.3 Observational Study

Twelve out of twenty-two subjects performed the experiments remotely, therefore, only ten people were observed during the experiment and gave input for the observational analysis. These are the most remarkable facts applicable to all tasks.

- The task is strongly visual: in general, users decide on relevance with just a quick look at the image thumbnail.
- Users feel more confident when searching in languages they know. More specifically, our Spanish users were reluctant to search in Dutch or German in spite of the relevance judgments being mostly visual.
- Remarkably, our users seemed to assume they could find everything in English. Only after some minutes searching they realized that this was not the case in our experiment.
- As the user acquires experience with the front-end, s/he uses the *assisted translation* search mode more often.
- The more experience the user has, the more s/he notices the mistranslations and search options. In general, however, the possibility of changing translations was largely unexploited.
- Most of the queries were submitted in Spanish. English (the most popular foreign language for our users) and Italian (which is very close to Spanish) were also used.
- We have identified three broad types of search strategies: i) a “depth first” strategy in which the user poses a general, broad query and then exhaustively inspects all the results; ii) a “breadth first” strategy, where the user makes many query variations and refinements, and makes only a quick inspection of the first retrieved results; iii) a “random” behavior, where the actions taken by the user can only be explained by “impulse”.

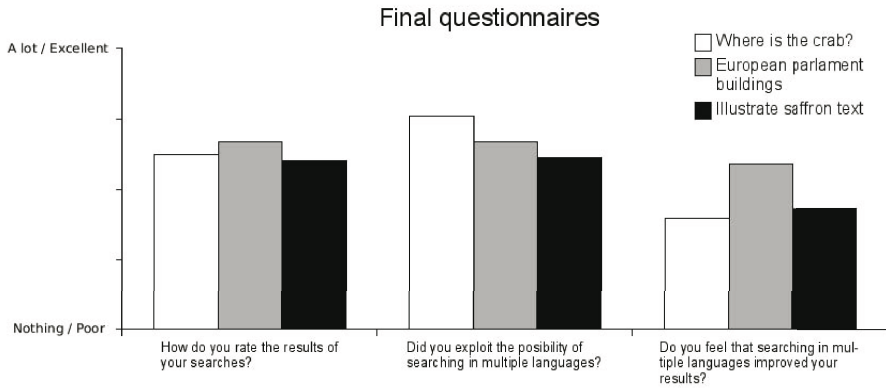


Fig. 4. Results of post-experiment questionnaire

3.4 User Questionnaires

Users were asked to fill in two different questionnaires. In the pre-search questionnaire, we asked users about their age and education, their previous experience searching images in the Web and using Flickr, and their language skills in the six languages proposed by our interface.

Two out of twenty-two users were 20-25 years old, twelve were 25-35 and eight were 35-45. Among them, there were four people with PhD studies, fourteen with different university degrees and four with high school diplomas. All of them were native Spanish speakers. Most people had some English skills and passive knowledge of other languages (mainly French and Italian), and most of them declared Dutch and German to be cryptic to them.

Lastly, after the experiment, users filled in an additional questionnaire about the perceived usefulness of the cross-language search facilities for every task and the degree of satisfaction with their performance.

Figure 4 sums up the data compiled in this last questionnaire. As shown in the first set of bars, most users were rather pleased with their overall performance, with slight differences among tasks. Then, we can also see that the more precision-oriented the task was, the more the users needed cross-lingual mechanisms. Lastly, for the “crab” and the “saffron” tasks, most subjects declared translation facilities did not help improve their results: in the first case, some users felt frustrated and they might have underestimated the usefulness of the multilingual search; in the second one, being an open-ended task it was not strictly necessary to search in other languages to complete a reasonable job.

4 Conclusions

In this paper, we presented the participation of UNED in the CLEF 2006 interactive task. Our goal was to measure the attitude of users towards cross-language

searching when the search system provides the possibility (as an option) of searching cross-language, and when the search tasks can clearly benefit from searching in multiple languages.

Our results indicate that, even in the most favorable setting (the results are images that can be often interpreted as relevant without reading their descriptions, and the system can make translations in a transparent way to the user), users often avoid translating their query into unknown languages. On the other hand, the learning curve to use cross-language facilities was fast. At the end of the experience, most users were using multilingual translations of their queries, although they rarely interacted with the system to fix an incorrect translation.

Acknowledgments

This work has been partially supported by the Spanish Government under project R2D2-Syembra (TIC2003-07158-C04-02), by the European Commission under project MultiMatch (FP6-2005-IST-5-033104) and the the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267). Javier Artiles and Víctor Peinado hold PhD grants by UNED (*Universidad Nacional de Educación a Distancia*).

References

1. Clough, P., Gonzalo, J., Karlgren, J.: iCLEF 2006 overview: searching the Flickr WWW photo-sharing repository. In: Working Notes of CLEF 2006 [In This volume] (2006)
2. López-Ostenero, F., Gonzalo, J., Verdejo, F.: UNED at iCLEF 2003: Searching Cross-Language Summaries. In: Comparative Evaluation of Multilingual Information Access Systems. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 450–461. Springer, Heidelberg (2004)
3. López-Ostenero, F., Gonzalo, J., Peinado, V., Verdejo, F.: Interactive Cross-Language Question Answering: Searching Passages versus Searching Documents. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 323–333. Springer, Heidelberg (2005)
4. Peinado, V., López-Ostenero, F., Gonzalo, J., Verdejo, F.: UNED at iCLEF 2005: Automatic highlighting of potential answers. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)

Providing Multilingual Access to FLICKR for Arabic Users

Paul Clough¹, Azzah Al-Maskari¹, and Kareem Darwish²

¹ Sheffield University, Sheffield, UK

² IBM, Cairo, Egypt

p.d.clough@sheffield.ac.uk

Abstract. In this paper we describe our submission for iCLEF2006: an interface that allows users to search FLICKR in Arabic for images with captions in a range of languages. We report and discuss the results gained from a user experiment in accordance with directives given by iCLEF, including an analysis of the success of search tasks. To enable the searching of multilingual image annotations we use English as an interlingua. An Arabic-English dictionary is used for initial query translation, and then Babelfish is used to translate between English and French, German, Italian, Dutch and Spanish. Users are able to modify the English version of the query if they have the necessary language skills to do so. We have chosen to experiment with Arabic retrieval from FLICKR due to the growing numbers of online Middle Eastern users, the limited numbers of interactive Arabic user studies for cross-language IR to date, and the availability of resources to undertake a user study.

1 Introduction

FLICKR¹ is a large-scale, web-based image database based on a large social network of online users. The application is used to manage and share personal (and increasingly more commercial) photographs. Currently FLICKR contains over five million accessible images, which are freely available via the web and are updated daily by a large number of users. The photos have multilingual annotations generated by authors using freely-chosen keywords (known as a folksonomy). Similar systems are also emerging for collections of personal videos² (e.g. youtube.com and CastPost).

Despite the popularity of FLICKR on a global basis, to-date there has been little empirical investigation regarding multilingual access to FLICKR. A common remark of Cross Language Information Retrieval (CLIR) is why would users want to retrieve documents that they (presumably) cannot read. Of course, in the case of image retrieval the motivation for CLIR is much stronger, because for many search tasks users are able to judge the relevance of images without the

¹ <http://flickr.com>

² <http://www.techcrunch.com/2005/11/06/the-flickr-of-video/>

need of additional text, thereby eliminating the need for translation of search results. This linguistic-neutrality of images makes text-based image retrieval an ideal application for CLIR.

For our submission to iCLEF 2006, we wanted to experiment with providing an interface which would enable users to query FLICKR in Arabic. We selected Arabic because of the availability of resources to us locally, the limited number of interactive Arabic user studies so far in CLIR, the growing number of online Middle Eastern users and the limited availability of online material in Arabic. According to Reuters and ABC Science Online³ there are currently only 100 million Arabic web pages, constituting 0.2% of the total pages on the web. There is no doubt that a system designed for Arabic users that is able to search English documents and other languages would open new possibilities both in terms of the quantity of accessible topics and the quality of the items retrieved.

The remainder of the paper describes the system developed for Arabic users, the experiments, results and conclusions. The three main aims of our work were: (1) to analyse the tasks offered by iCLEF, (2) to analyse our initial interface: query translation and use of English as an interlingua, and (3) to observe the searching behaviour of users.

2 The System: FLICKRArabic

We developed an Ajax-based online application called FLICKRArabic (shown in Fig. 1) to provide query translation to the FLICKR API⁴. The system centres on translating user's queries from Arabic into English (the interlingua) and then consequently translating into French, Spanish, German, Italian or Dutch. This is necessary because many translation resources provide only Arabic to English translation. The English translation is shown to users who can modify the query if they have sufficient language skills. In the case of polysemous Arabic query words, translations for all available senses are displayed to the user and the user is able to ignore or select the correct translations. Translation between Arabic-English is performed using a bilingual dictionary (described in Section 2.1). Translation between English and other languages is performed by using a custom-built wrapper for Babelfish⁵, which is an online Machine Translation (MT) system. This means that beyond English, users have little control over translation. The design and implementation is described further in Section 2.2.

2.1 Language Resources

Two translation resources have been used to create the application. The first is an Arabic-English bilingual dictionary, the second is the online MT tool Babelfish. To create the bilingual dictionary, two bilingual term lists were constructed

³ <http://www.abc.net.au/news/newsitems/200604/s1624108.htm>

⁴ <http://www.flickr.com/services/api/>

⁵ <http://babelfish.altavista.com>

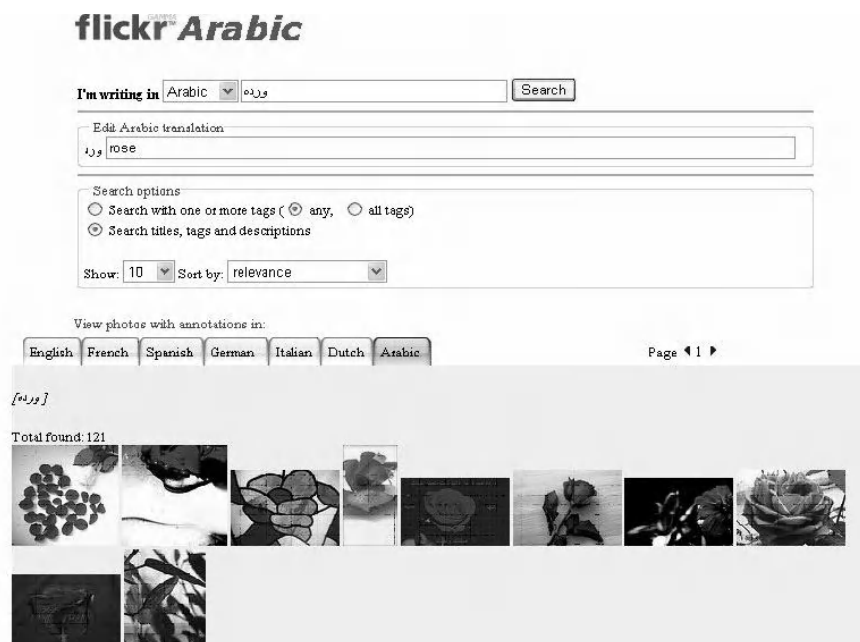


Fig. 1. FLICKRArabic interface

using two Web-based MT systems, namely Tarjim⁶ and Al-Misbar⁷. In each case, a set of unique isolated English words found in a 200 MB collection of Los Angeles Times news stories was submitted for translation from English into Arabic [4]. Each system returned at most one translation for each submitted word. In total, the combined bilingual term lists contained 225,057 unique entries. In preprocessing the Arabic text, all diacritics and kashidas (character elongations) were removed, the letters *ya* and *alef magsoura* were normalised to *ya* and all the variants of *alef* and *hamza*, namely *alef*, *alef hamza*, *alef maad*, *hamza*, *waw hamza*, and *ya hamza*, normalised to *alef*, and lastly all words were stemmed using Al-stem⁸.

2.2 Interface Design and Functionality

Design of the system has emerged from an interactive evaluation design process. A user-centered approach was implemented, where five Arabic potential users were involved during the design and implementation phases of the system (following the advice of [7]). During the pilot session, users were observed and questioned about their cross-language actions (e.g. editing the translation of the

⁶ <http://tarjim.ajeeb.com>, Sakhr Technologies, Cairo, Egypt.

⁷ <http://www.almisbar.com>, ATA Software Technology Limited, North Brentford Middlesex, UK.

⁸ <http://www.glue.umd.edu/~kareem/research/>

Arabic query and flipping through the results of other languages). Previous research which has considered user interaction in the formulation of multilingual queries includes the Keizai system [9], ARCTOS [10], MULINEX [2], WTB [8], MIRACLE [5], EUROVISION [3] and CLARITY [11]. The basic functionality of this system is as follows:

- Users can search FLICKR using initial English or Arabic queries
- If searching in English, the system calls the FLICKR API and displays results
- If searching in Arabic:
 - The query is first converted from UTF8 to CP1256 (using *iconv -f utf8 -t cp1256*) and stemmed (using the *stem_cp1256.pl* Perl program)
 - The query is then translated into English using the Arabic-English dictionary
 - The system returns all dictionary matches (senses and synonyms) for each translated query
 - Users can modify the English translation by deleting unwanted terms or adding their own. Previous research has shown it useful to enable this for most bilingual users
 - The English query is then used for searching FLICKR using the API⁹
 - The user can additionally view photos annotated in Arabic only (for comparison with other languages)
- Results for each language are displayed in separate tabs with the total number of images found displayed (when each language selected)
- The user can view photos with annotations in any one of the five languages: French, Spanish, German, Italian and Dutch¹⁰
- Users can select the following search options:
 - Display 10, 20, 50, 100 or 200 images per page
 - Sort images by a relevance score or interestingness
 - Search all annotation text (titles, tags and descriptions) or tags only (all query terms or any)

Results from an English search were displayed to users first (left-hand tab), because during initial tests most users were able to make use of images with English annotations (ordering of languages was therefore arbitrary). Results for Arabic were also provided to enable users to compare results from searching FLICKR using purely Arabic.

3 The Experiment

To obtain feedback on the implemented system, we recruited 11 native Arabic speakers to carry out the tasks specified by iCLEF [6]. The subjects were undergraduate and postgraduate students with a good command of English¹¹. The mean age of the 11 users was 28 years old, and 85% stated they typically

⁹ Only photos posted before 1/6/2006 were returned.

¹⁰ If the query is not translated, it remains in English and results for this are returned to the user.

¹¹ We are currently planning experiments with monoglot users: those who can only make use of Arabic.

searched the Web using English. They also had the following characteristics: 82% used the Internet several times a week, all had a great deal of experience with point-and-click interfaces, 46% searched for images very often and 46% of those people often found what they were searching for.

Subjects were asked to perform 3 tasks: (1) a classical ad-hoc task: “find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside” (parliament); (2) a creative instance-finding task: “find five pictures to illustrate the text - the story of saffron - with the goal being to find five distinct instances of information described in the narrative: saffron, flower, saffron thread, picking the thread/flower, powder, dishes with saffron (saffron); and (3) a visually orientated or known-item task: given a picture, find the name of the beach on which the crab is resting (crab). More details of the tasks can be found in [6].

In these experiments, users first completed a preliminary questionnaire, then spent 20 minutes on each task. Tasks were assigned randomly to users to reduce the effects of task bias on the results (e.g. user 1 performed task 2, 3, then 1, user 2 performed task 3, 1, then 2). We also asked subjects to perform a final search where they were able to search for images on any topic. Finally, we asked users to complete a questionnaire to establish their overall satisfaction with and impressions of the system. During the experiment, we recorded some attributes of the task such as time taken and queries input, as well as taking notes of the user’s searching behaviour during each task.

4 Results and Observations

4.1 User Effectiveness

We first discuss how well users were able to perform the tasks¹². Almost all users (10 out of 11) were able to perform task 3 (crab) successfully¹³. Table 1 shows the results for task 1 (parliament). For this task, from the total number of images found, we deemed which ones are *correct* and of these we counted the number of *unique* European parliament buildings. To compute recall we divided unique by correct; for precision we computed correct divided by total. We also divided the total number of pictures found between those of the inside of the building versus the outside. Across all users we obtained a recall of 0.69 and precision of 0.84. This varied between users with some scoring higher recall (e.g. user 3) and others achieving higher precision (e.g. users 7 and 9). Of the images found, almost twice as many were of the outside of buildings.

Table 2 shows the results for task 2 (saffron). In this task users were asked to find 5 images to illustrate each part (or instance) to the story of saffron. Users were given one point for retrieving each instance (*counted*) and no credit

¹² The system effectiveness and the correlation between user and system effectiveness is explored further in [1].

¹³ Despite users not being familiar with German, they were able to recognise the name of the beach.

Table 1. Search results for task 1 (parliament)

User	total	found	correct	unique	recall	precision	inside	outside
1*	—	—	—	—	—	—	—	—
2	20	11	3	0.27	0.55	9	11	
3	11	8	7	0.88	0.73	4	7	
4	20	17	11	0.65	0.73	4	13	
5	13	12	8	0.67	0.92	4	9	
6	12	10	9	0.90	0.83	3	9	
7	12	12	8	0.67	1.00	5	9	
8	8	7	6	0.86	0.88	5	3	
9	12	12	9	0.75	1.00	0	12	
10	13	10	7	0.70	0.77	4	9	
11	10	9	7	0.78	0.90	3	7	
Total	131	108	75	0.69	0.84	41	80	

*The system did not function correctly for this user during this task

for repeated instances (i.e. no additional credit for selecting two images of the same aspect of the story). Results show the number of images found for each aspect/instance and the precision (counted divided by total). Overall precision was 0.70 for this task and again, precision varied between users as some were good at instance-finding (e.g. users 3 and 9) and others were less successful (e.g. users 5 and 11).

Table 2. Search results for task 2 (saffron)

User	flower	thread	food	powder	picking	total	counted	precision
1	3	2	—	—	—	5	2	0.40
2	1	1	2	—	1	5	4	0.80
3	1	1	1	1	—	4	4	1.00
4	1	2	2	—	—	5	3	0.60
5	3	—	—	—	—	3	1	0.33
6	2	1	1	—	1	5	4	0.80
7	1	1	1	1	1	5	5	1.00
8	2	1	1	—	1	4	5	0.80
9	1	1	1	1	1	5	5	1.00
10	—	4	1	—	—	5	2	0.40
11	2	1	2	—	—	5	3	0.60
Total	18	17	11	2	4			0.70

Table 3 shows the number of users who judged results for each language and task as highly relevant, partially relevant or not relevant. It would appear that users found relevant images with annotations in most of the languages, except Arabic. Users commented that they were disappointed with the Arabic results. This would suggest that multilingual access to FLICKR could improve retrieval. For task 2 (saffron), most users found relevant images in Italian which is likely due to the narrative mentioning Italy.

Table 4 shows the user’s satisfaction with the accuracy and coverage of the search results. Overall it appears that users were very satisfied with the accuracy of search results from FLICKRArabic for tasks 1 and 3, but less satisfied with the accuracy of results for task 2. Similarly, users appear in general satisfied with the coverage of results (again less so for task 2). We asked users whether

Table 3. User relevance per language

Language	Task 1 (parliament)			Task 2 (saffron)			Task 3 (crab)		
	Highly	Partially	Not	Highly	Partially	Not	Highly	Partially	Not
English	9	2	0	7	4	0	3	5	3
French	4	5	2	7	4	0	2	7	2
Spanish	5	4	2	7	4	0	4	6	1
German	4	5	2	8	3	0	2	6	3
Italian	5	4	2	6	5	0	3	3	5
Dutch	2	5	4	3	5	3	3	3	5
Arabic	0	0	11	0	0	11	5	3	3

they felt accuracy or coverage was more important for the tasks 1 and 2. In task 1 (parliament), 7 users favoured accuracy and 4 preferred coverage; in task 2 (saffron) 8 users favoured accuracy and 3 preferred coverage. Overall, it would appear that users would prefer more accurate results, which likely reflects the precision-orientated nature of the tasks.

Table 4. User’s satisfaction with accuracy and coverage

Task	Accuracy			Coverage		
	Highly	Partially	Not	Highly	Partially	Not
1 (parliament)	10	1	0	8	2	1
2 (saffron)	5	6	0	5	6	0
3 (crab)	10	1	0	9	2	0

We asked users about the usefulness of the results for each task and overall 70% of users were very satisfied with the results (30% partially satisfied). Table 5 indicates how user rate the importance of factors in helping them to determine the usefulness of the images for each search task. Users rated these as very important (v. imp.), important (imp.) and not important (unimp). For task 1, users found textual information very important in addition to the image itself. We expected this as users need to check the annotations to determine whether a parliament building is European or not. In task 2, users found the image and caption to be the most important for determining relevance. Users were able to identify possible pictures of saffron, but needed the captions to confirm their decision. In task 3, as expected, users found the visual content of the photos most useful. This reflects the fact that this task is more visual in nature. Users also found foreground and background text useful to determine the beach where the crab was placed.

4.2 Users’ Search Behaviour

The following observations regarding search behaviour were observed during the experiment: users typically viewed initial results in English before trying other languages. This is because they were able to read annotations in English. Two main strategies for searching prevailed: some users input fewer queries and looked through many pages of results; others input many queries and if no relevant found

Table 5. User-rated usefulness of image attributes

	Task 1 (parliament)				Task 2 (saffron)				Task 3 (crab)			
	V. imp.	Imp.	Unimp.		V. imp.	Imp.	Unimp.		V. imp.	Imp.	Unimp.	
Image only	8	3	0	8	2	1	11		0	0		
Image and caption	10	0	1	8	3	0	5		6	0		
Comments	7	3	1	5	5	1	1		6	4		
Foreground details	7	1	3	4	5	2	6		3	2		
Background details	1	3	7	1	3	7	6		4	1		
Previous knowledge	2	2	7	5	4	2	2		1	8		

in the first page of results, they reformulated the query. For the search results, some users would systematically look through results for each language from left to right; others would start with the languages which returned the least number of results (testing each language first). Most users selected 100 images at a time to view in the search results suggesting they are able (and willing) to view a large number of thumbnails.

4.3 Users’ Comments on the Tasks

To determine the success of each task, we gathered users’ comments on different aspects of the tasks as shown in Tables 6 and 7. Overall (from Table 6) it would appear that tasks 1 and 3 were the clearest, with task 1 being the easiest, task 3 being the most familiar, and tasks 2 and 3 being the most interesting to users. Interestingly, the majority of users did not find any of the tasks relevant to them. This is primarily because these topics were not designed specifically for Arabic users who are likely to search for different topics than Europeans and the tasks themselves were not entirely realistic (e.g. searching for a crab on a beach to find a specific location). It was also interesting to find that users were reluctant to search for crab, because the equivalent Arabic word has another sense, namely cancer. Also, the query for saffron was alien to most of the male searchers who did not cook and therefore were unsure what saffron was or looked like.

Table 6. User’s assessment of the search tasks (1)

	Task 1 (parliament)				Task 2 (saffron)				Task 3 (crab)			
	Highly	Partially	Not		Highly	Partially	Not		Highly	Partially	Not	
Clear	9	1	1	4	4	3	11		0		0	
Easy	8	3	0	7	4	0	6		4		1	
Familiar	7	3	1	3	8	0	9		2		2	
Interesting	5	3	2	6	5	0	7		3		1	
Relevant	2	4	5	1	4	6	1		4		6	

In Table 7 that compares between tasks, users found task 1 to be the most interesting and easiest, tasks 1 and 3 to be the most enjoyable and task 1 to be the most realistic out of the three tasks. Users commented that task 3 was very unrealistic and did not represent the type of search task they would perform. For iCLEF organisers, this might indicate that concentrating on adhoc search is more likely to represent users’ tasks and is less likely to be artificial.

Table 7. User’s assessment of the search tasks (2)

Task	Interesting			Easiest			Enjoyable			Realistic	
	Most	Somewhat	Least	Most	Somewhat	Least	Most	Somewhat	Least	Yes	No
Task 1 6	1		4	7	1	3	4	5	2	6	5
Task 2 3	6		2	1	7	3	2	6	3	3	8
Task 3 2	4		5	3	4	4	4	3	4	1	10

4.4 Free-Search Task

If users could search for their own topics, what would they search for? We asked users to submit their own queries to determine the types of topics that would be representative or interesting to this user group. Table 8 shows the queries submitted by each user, the query language used, and the language of the annotations viewed in the results. As expected, most users typed queries more related to their culture (e.g. names of places such as Oman and Leptis and objects such as a mosque) and interests (e.g. welding). Most users searched using Arabic but found results in other languages helpful or useful. For queries containing out-of-vocabulary terms (e.g. Leptis), users had sufficient language skills to search in English.

Table 8. User’s queries for free-search task

User	Query	Language of query	Languages of viewed results
1	Damascus	Arabic	Arabic, German, then English
2	Leptis (a city in Libya)	Arabic (not in dict), then English	Arabic, then English
3	Shef Uni, Mecca, Jeddah	Arabic	English, Arabic, then right-left
4	Tower, Skyscraper, Suspension bridge	Arabic	Arabic, English, then left-right
5	Damascus, Mecca	Arabic	Arabic
6	Orientalism	Arabic (not in dict), then English	English
7	Welding, Pyramid	English	English
8	Mosque	Arabic	English, German, Italian, Dutch, Arabic
9	Castles in Oman, andalus, Hamraa palace	Arabic	Arabic, then English (few results)
10	Cats, Muscat	Arabic and English	English, Arabic, French; English, French, left-right
11	Manchester, England, Libya	English	English

4.5 Overall User Comments

Table 9 indicates overall users’ comments about the system we implemented. This represents users’ satisfaction as recorded by the categories: very satisfied, partially satisfied and not satisfied. User’s were very positive about the system and definitely found the provision of multilingual access to be useful to them (the ability to view pictures with annotations in various languages). Most users (9 out of 11) were very satisfied with the multilingual search results (compared against using FLICKR as is). However, the majority of these users were not happy with the query translation (7 partially satisfied and 1 not satisfied). Indeed we found

that because this user group had good English language skills, the translation from Arabic to English was actually an unnecessary step and most preferred to formulate and modify queries in English (10 out of 11 users were willing to modify the English version of the query, and all users would enter the English version of a query term if not in the dictionary and add synonyms).

Table 9. User’s overall satisfaction rating of the system

	Very	Partially	Not
Overall success	7	4	0
Multilingual usefulness	11	0	0
Multilingual satisfaction	9	2	0
Use system again	8	0	3
Recommend system to friend	11	0	0
Easiness of use	11	0	0
Quality of translation	3	7	1
Willingness to modify query	10	0	1
Entry of synonyms	11	0	0

Most users (8 out of 11) would use the system again and all users would recommend the system to a friend or colleague. All users were very satisfied with the ease with which the system could be used. Many users said they would have viewed most non-English annotations if translations had been provided in English (or Arabic). This suggests that some form of document translation could improve a users’ search experience.

5 Discussion and Conclusions

The goal of our work was to build and test a simple Arabic query interface to FLICKR enabling users to view images with annotations in a range of languages. To enable Arabic translation into multiple languages, we first translated into English (interlingua) using a bilingual Arabic-English dictionary. From initial user testing, we decided to show users the English translations and allow them to edit as desired. The English version of the query was then translated into other languages as users requested to view results in those languages using the Babelfish MT system.

Overall with the group of users recruited for this experiment, we found that providing Arabic as an initial query language was unnecessary and caused more frustration than usefulness due to poor translation or out of dictionary words. Users were much happier submitting and reformulating queries in English (particularly for the tasks set which were orientated to Europeans as opposed to Arabs). Some users expressed the need for the ability to search in Arabic in some cases (e.g. when they are unable to formulate a query in English), but this was not the case for most of these tasks. However, users did comment that being able to start the search in Arabic to obtain some terms in English was a useful way to begin their search. Some users also suggested that being able to combine Arabic and English queries would be useful.

Compared to the current FLICKR system, it would seem that being able to submit an English query and translate it into multiple languages is considered as very beneficial to end users. It was particularly apparent with some queries where search results are very much language-dependent and different (e.g. searching for *car* typically produces British-built cars for English and *voiture* produces French-built cars). It would, therefore, seem more important to focus on this part of the system than initial query translation. Users were generally complementary of our system and they were able to carry out the search tasks set with reasonable success: overall precision of 0.84 for task 1, precision of 0.70 for task 2 and 10 out of 11 users completed task 3 successfully. Further work is planned in the following areas:

- Running the experiment again with users who are less proficient in English and would be less likely to reformulate English versions of queries.
- Improving query translation by increasing the size of the dictionary, handling the translation of phrases, and enabling the user to correct erroneous dictionary entries.
- Presenting results to users with different ranking and clustering strategies to reduce the number of images users view to find relevant images.
- Performing simultaneous searches in different languages to provide users with a summary of the number of results in each language. This might help users who typically view languages which exhibit the fewest number of results first.
- Translating image annotations in FLICKR results.

Acknowledgments

This work has been partially supported by the MultiMatch project (FP6-2005-IST-5-033104) under the EU Sixth Framework.

References

1. Al-Maskari, A., Clough, P., Sanderson, M.: User's effectiveness and satisfaction for image retrieval. In: Workshop Information Retrieval, University of Hildesheim, Germany (October 9–11, 2006)
2. Capstick, J., Diagne, A., Erbach, G., Uszkoreit, H., Cagno, F., Gadaleta, G., Hernandez, J., Korte, R., Leisenberg, A., Leisenberg, M., Christ, O.: Mulinex: Multilingual web search and navigation (1998)
3. Clough, P., Sanderson, M.: User experiments with the eurovision cross-language image retrieval system. 57(5), 679–708 (2006)
4. Darwish, K., Oard, D.W.: Clir experiments at maryland for trec 2002: Evidence combination for arabic-english retrieval (2002)
5. Dorr, B.J., He, D., Luo, J., Oard, D.W., Schwartz, R., Wang, J., Zajic, D.: iclef 2003 at maryland: Translation selection and document selection. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 435–449. Springer, Heidelberg (2004)

6. Gonzalo, J., Karlgren, J., Clough, P.: iclef 2006 overview: Searching the flickr www photo-sharing repository. In: Proceedings of Cross-Language Evaluation Forum (CLEF2006) Workshop, Alicante, Spain (2006)
7. Hackos, J.T., Redish, J.C.: User and task analysis for interface design. John Wiley & Sons, Inc., New York (1998)
8. Penas, A., Verdejo, F., Gonzalo, J.: Website term browser: Overcoming language barriers in text retrieval. *Journal of Intelligent and Fuzzy Systems* 12(3–4), 221–233 (2002)
9. Ogden, W., Cowie, J.: Keizai: An interactive cross-language text retrieval system(1999)
10. Ogden, W., Davis, M.W.: Improving cross-language text retrieval with human interactions. In: Proceedings of the Hawaii International Conference on System Science (HICSS-33), vol. 3 (2000)
11. Petrelli, D., Levin, S., Beaulieu, M., Sanderson, M.: Which user interaction for cross-language information retrieval? design issues and reflections. *Journal of the American Society for Information Science and Technology (JASIST)* 57(5), 709–722 (2006)

Trusting the Results in Cross-Lingual Keyword-Based Image Retrieval*

Jussi Karlgren and Fredrik Olsson

The Swedish Institute of Computer Science, Stockholm, Sweden
jussi@sics.se

Abstract. This paper gives a brief description of the starting points for the experiments the SICS team has performed in the 2006 interactive CLEF campaign.

1 Keywords Can Be Obscure

Flickr, the database chosen for this years interactive CLEF experiments, does not offer direct image search capabilities; searching for images in the database is done by searching keywords attached to the images by photographers or in some cases other viewers.

Searching for images using textual cues is a challenging task. There are many non-conventionalised keyword schemes, especially in a context such as Flickr, where keywords can be freely chosen. In Flickr, to compound the complexity of the task, the keywords frequently are given in several languages.

This means that we expect image viewers to be unsure of finding the right answer in Flickr.

Hypothesis 1: Viewers are uncertain of whether they have found everything that they set out to find.

2 Terms Are Related by Their Distribution

How could technology improve the sense of confidence viewers have in the results they have found? Our contention is that the descriptive qualities of a term are partially obscure to a user, and that this obscurity is exacerbated by cross-lingual effects - trying to use keywords in a foreign language will be more daunting than using terms in one's own. From other recent research we have investigated the effects of using distributional data to postulate semantic relations between terms [3] and in this experiment we make use of distributional similarity to provide related terms. We display the terms used by the user, together with distributionally related terms in a word space screen for the user.

* This work has been partially supported by the European Commission, through the DELOS Network of Excellence (FP6 G038-507618).

Hypothesis 2: By displaying the terminological space of tags related by distributional analysis to the ones used by the viewer we can encourage users to further search activity.

3 Relevance Can be Imprecise

Evaluation of user satisfaction is a contentious issue. The target notion of “Relevance” - related to and inspired by the everyday notion - in evaluating information retrieval systems is formalized to be an effective tool for focused research. Much of the success of information retrieval as a research field is owed to this formalization. But “Relevance” does not take user satisfaction or information need fulfilment into account. It is unclear how it could be generalized to the process of retrieving other than factual accounts. It is binary, where the intuitive and everyday understanding of relevance quite naturally is a gliding judgment. It does not take sequence of presentation into account - after seeing some information item, others may immediately lose relevance.

Trying to extend the scope of an information retrieval system so that it is more task-effective, more personalized, or more enjoyable will practically always carry an attendant cost in terms of lowered formal precision and recall as measured by relevance judgments. This cost is not necessarily one that will be noticed, and most likely does not even mirror a deterioration in real terms - it may quite often be an artefact of the measurement metric itself. Instead of being the intellectually satisfying measure which ties together the disparate and vague notions of user satisfaction, pertinence to task, and system performance, “Relevance” gets in the way of delivering all three.

The underlying hypothesis of our research activities is that the target notion for evaluating information access systems needs to cater for more than just topical relevance.

Hypothesis 3: User satisfaction is measurable and quantifiable in some respects.

4 First Experiment

Twelve users first performed the three experimental tasks (as described in the track overview[1]):

- Topical ad-hoc retrieval: *Find as many European parliament buildings as possible, pictures from the assembly hall as well as from the outside.*
- Creative open-ended retrieval: *Find five illustrations to the article “The story of saffron”*
- Visually oriented task: *What is the name of the beach where this crab is resting?* (along with a picture of a crab lying in the sand).

This year’s iCLEF did not contribute to the evaluation by requiring any specific metrics to be employed. Our evaluation has centered on the user experience of completion and satisfaction rather than accuracy. The three metrics we have employed are

1. Happy (all tasks): Are you satisfied with how you performed the task?
2. Complete (creative task, ad-hoc task): Did you find enough, or would you have continued if there had not been a time limit? How long would it have taken to make you satisfied, do you think?
3. Quality (creative task): Compare the illustrations with some given set. Are any of these better than your retrieved images?

As an additional metric, after a query has been performed, a terminology display of the terms used is displayed, where the terms employed by the user are displayed in order, together with other terms, related to the original terms by distributional metrics. The users can browse through the word space, and return to Flickr to rerun searches: we tabulate the number of additional searches made through this interface.

4.1 First Results and Lessons Learnt

Firstly, our test subjects expressed hesitation to use terminological suggestions which they did not themselves understand, even if they would have been likely to understand the results of the retrieval. This gave us further reason to think about trust and what mechanisms would motivate users to explore information sources in any given direction.

Secondly, the visual task proved either very difficult or very simple: if the subjects figured out they were to use the German term for “crab” they immediately solved the task - without that insight they did not succeed. The results of that task were too binary to provide any deeper insights in task solution strategies or suitable aid systems.

Thirdly, the creative task proved both challenging and entertaining, but takes some further finesse to evaluate appropriately. We chose to explore the constraints of that task further in a continued experiment.

5 Second Experiment: Simulated Social Context

One of the basic human behavioural guides is that other people’s actions are useful guides to find interesting and effective places to be. Most people like to hang out with their crowd, or indeed with any crowd. This is well established both anecdotally and research-wise, and we take it as given. The research field of social navigation is to a large extent based on this hypothesis[2]. In addition, there is a complementary and not entirely unrelated hypothesis that accounts for much of social behaviour which we study as part of this experiment: people like to make a mark on the world. This is less self-evident, and will involve much larger variation between people - personality differences and social effects of various types dampen or strengthen the tendency for people to express themselves. But even here, we will for the purposes of this study presume that people in general wish to make a noticeable difference in some way.

We want to see if the above two behavioural motivating factors will translate into observable behaviour in interacting with computer systems. Specifically, we believe that

in an assessment task - a task where people assess the suitability of some item for some purpose - these factors will effect how people grade and judge items that they are informed others have assessed before them.

Hypothesis 4: People will tend to follow their peers in assessing subjective worth of items.

Hypothesis 5: People like to give assessments that make a difference for the end value of the item under consideration.

These hypothesized mechanisms are to a degree mutually incompatible. We will try study them simultaneously in the same experimental setup. We used the given sets of illustrations as a target category for our experiment: the assessment of images is for a nonverbal and impressionistic task, which we presume will aid us in coaxing out the effects of social context from the results of the experiment. In this case, the subjects were asked to assess the illustrations given by the subjects of the previous experiment.

Any social interface hinges crucially on the selection of the peer group to relate behaviour to - in this case, the selection of users is done from an established social context (visitors to a personal photo blog) but this fact is not stressed in the interface itself. The interface displays viewing statistics, and the result of the study will be based on observed difference in behaviour dependent on different displayed statistics, following the hypothesis that viewing statistics will both influence subjects to want to follow the example of others as well as to want to issue assessments that make a difference to the end result.

Along with the image sets, statistics about how many had viewed the various sets as well as their average rank were displayed. Unknowingly to the subjects, these figures were all fake and were given in two versions. Figure 1 gives the numbers that were displayed in the two conditions: in condition (B) the “previously viewed by” figure given to the subjects was higher than the other. Two of the sets were not changed, to preserve a control case.

Set Previously viewed by Score			
	A	B	
1	98	196	4.02
2	120	120	3.11
3	113	226	2.09
4	94	188	4.10
5	181	181	3.20

Fig. 1. Data for the simulated social context

5.1 Continued Results

The results from our first survey turned out to go in several directions. 35 sub-jects completed the survey. A summary of the results can be found in Figure 2.

Set	Score	
	A	B
1	3.37	3.88
2	3.52	2.75
3	4.47	4.56
4	2.21	1.94
5	2.47	1.88

Fig. 2. First results from the study

5.2 Hypotheses Revisited - A Disappointment

Our expectation was, following our hypotheses, that in the condition with large numbers of previous views, subjects would tend to adhere to the displayed score - following the crowd, as it were; in the conditions with fewer views they would tend to diverge - finding an opportunity to make a difference.

The differences for sets 1 and 5 are the most significant ($0.9 < p < 0.95$ by Mann Whitney U) but the results are equivocal. In the case of set 1, indeed the subjects who were given the larger number of previous viewers did score the set closer to the given score. In the case of set 5, the given figures for the sets were identical, and only indirectly can the displayed figures for previous viewers have influenced the result. The varying tendency to diverge from the given score cannot be reliably traced to the two different conditions.

5.3 Lessons Learnt

Better control for content: In the case of section 3, apparently the quality of the selected images so overwhelmingly convinced the subjects as to maximise any other effects.

Better distinction between conditions: In our next cycle of the study we will boost the differences between the two conditions to better tease out differences.

Interaction design issues: We attempted to design an interface where the variables under study would be prominent without being obtrusive and an interaction case which would at least resemble a real application. The current interface enjoined subjects to make decisions even where they may not have a strong opinion. We will amend this in the next cycle.

6 Conclusions

We found that the experimental methodology was promising, the subjects interested, opinionated, and committed to providing results of some usefulness. The results, by contrast, were somewhat equivocal. We will pursue this design further in continued interactive experiments.

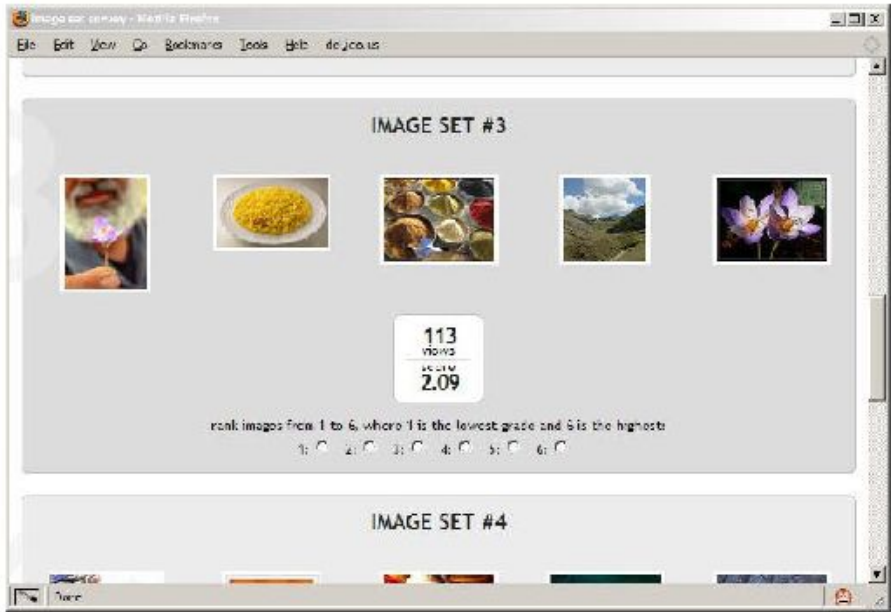


Fig. 3. A series of images serving as an answer to the creative task

References

1. Gonzalo, J., Karlgren, J., Clough, P.: iclef 2006 overview: Searching the flickr www photo-sharing repository. [In This volume] (2006)
2. H"o"ok, K., Munro, A., Benyon, D.: Designing Information Spaces: The Social Navigation Approach. Springer, Heidelberg (2002)
3. Karlgren, J., Sahlgren, M.: Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering* 11(3), 327–341 (2005)

Overview of the CLEF 2006 Multilingual Question Answering Track

Bernardo Magnini¹, Danilo Giampiccolo², Pamela Forner², Christelle Ayache³,
Valentin Jijkoun⁴, Petya Osenova⁵, Anselmo Peñas⁶, Paulo Rocha⁷,
Bogdan Sacaleanu⁸, and Richard Sutcliffe⁹

¹ FBK-irst, Trento, Italy
magnini@itc.it

² CELCT, Trento, Italy
{giampiccolo, forner}@celct.it

³ ELDA/ELRA, Paris, France
ayache@elda.fr

⁴ Informatics Institute, University of Amsterdam, The Netherlands
jijkoun@science.uva.nl

⁵ BTB, Bulgaria

petya@bultreebank.org

⁶ Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain
anselmo@lsi.uned.es

⁷ Liguatca, SINTEF ICT, Norway and Portugal
Paulo.Rocha@alfa.di.uminho.pt

⁸ DFKI, Germany
Bogdan.Sacaleanu@dfki.de

⁹ DLTG, University of Limerick, Ireland
richard.sutcliffe@ul.ie

Abstract. Having being proposed for the fourth time, the QA at CLEF track has confirmed a still raising interest from the research community, recording a constant increase both in the number of participants and submissions. In 2006, two pilot tasks, WiQA and AVE, were proposed beside the main tasks, representing two promising experiments for the future of QA. Also in the main task some significant innovations were introduced, namely list questions and requiring text snippet(s) to support the exact answers. Although this had an impact on the work load of the organizers both to prepare the question sets and especially to evaluate the submitted runs, it had no significant influence on the performance of the systems, which registered a higher Best accuracy than in the previous campaign, both in monolingual and bilingual tasks. In this paper the preparation of the test set and the evaluation process are described, together with a detailed presentation of the results for each of the languages. The pilot tasks WiQA and AVE will be presented in dedicated articles.

1 Introduction

Inspired by previous TREC evaluation campaigns, QA tracks have been proposed at CLEF since 2003. During these years, the effort of the organisers has been focused on two main issues. One issue was to offer an evaluation exercise characterised by cross-linguality, covering as many languages as possible. From this perspective, major

attention has been given to European languages, adding at least one new language each year, but keeping the offer open to languages from all-over the world, as the use of Indonesian shows. The other important issue was to maintain a balance between the established procedure inherited from the TREC campaigns and innovation. This allowed newcomers to join the competition and, at the same time, offered “veterans” more challenges. Following these principles, in QA@CLEF 2006 two pilot tasks, namely WiQA and Answer Validation Exercise (AVE), were proposed together with a main task. As far as the latter is concerned, the most significant innovation was the introduction of LIST questions, which had also been considered for previous competitions, but had previously been avoided due to the problems that their selection and assessment implied.

Other important innovations consisted in the possibility to return more than one answer per question, and by the request to provide text snippets together with the docid to support the exact answer. All these changes implied the necessity of introducing new evaluation measures, which would account also for List and multiple answers. Nevertheless, the evaluation process proved to be more complicated than expected, partly because of the excessive workload that multiple answers represented for groups already in charge for a larger number of runs. As a consequence, some groups, like the Spanish and the English ones, could only correct one answer per question, which decreased the possibility of comparisons between runs.

As a general remark, it can be said that the positive trend in participation registered in the previous campaigns was confirmed, and 10 new participants joined the competition from Europe, Asia and America.

As reflected in the results, systems' performance improved considerably, with the Best Accuracy increasing from 64% to 68% in the monolingual tasks, and, more significantly, from 39% to 49% in the bilingual ones.

This paper describes the preparation process and presents the results of the QA track at CLEF 2006. In section 2, the task is described in detail. The different phases of the Gold Standard preparation are exposed in section 3. After a quick presentation of the participants in section 4, the evaluation procedure and the results are reported respectively in section 5 and 6. In section 7, some final considerations are given about this campaign and the future of QA@CLEF.

2 Tasks

In 2006 campaign, the procedure consolidated in previous competitions was used. Accordingly, there was a main task (which was comprehensive of a monolingual task and several cross-language sub-tasks), and two pilot tasks described below:

1. **WiQA:** developed by Maarten de Rijke. The purpose of the WiQA pilot is to see how IR and NLP techniques can be effectively used to help readers and authors of Wikipages access information spread throughout Wikipedia rather than stored locally on the pages.[2]
2. **Answer Validation Exercise (AVE):** A voluntary exercise to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by a QA system. The basic idea is that once

a pair [answer + snippet] is returned by a QA system, a hypothesis is built by turning the pair [question + answer] into the affirmative form. If the related text (a snippet or a document) semantically entails this hypothesis, then the answer is expected to be correct [3].

Two specific papers in the present Working Notes are dedicated to these pilot tasks. More detailed information, together with the results, can be found there.

In addition to the tasks proposed during the actual competition, a "time-constrained" QA exercise was proposed by the University of Alicante during the CLEF 2006 Workshop. In order to evaluate the ability of QA systems to retrieve answers in real time, the participants were given a time limit (e.g. one or two hours) in which to answer a set of questions. These question sets were different and smaller than those provided in the main task (e.g. 15-25 questions). The initiative was aimed towards providing a more realistic scenario for a QA exercise.

The main task was basically the same as in previous campaigns. Some new ideas were implemented in order to make the competition more challenging. The participating systems were fed a set of 200 questions, which could be about:

- facts or events (F-actoid questions);
- definitions of people, things or organisations (D-efinition questions);
- lists of people, objects or data (L-ist questions).

The systems were then asked to return from one to ten exact answers. "Exact" meant that neither more nor less than the information required is given. The answer needed to be supported by the docid of the document(s) in which the exact answer was found, and by one to ten text snippets which gave the actual context of it.

The text snippets were to be put one next to the other, separated by a tab. The snippets were substrings of the specified documents. They should provide enough context to justify the exact answer suggested. Snippets for a given response had to be a set of sentences of not more than 500 bytes in total (although for example the Portuguese group accepted – and actually preferred – length to be specified in sentences). There were no particular restrictions on the length of an answer-string, but unnecessary pieces of information were penalized, since the answer was marked as *ineXact*. Since Definition questions may have long strings as answers, they were (subjectively) assessed mainly on their informativity and usefulness, and not on exactness. The tasks were both:

- monolingual, where the language of the question (Source language) and the language of the news collection (Target language) were the same;
- cross-lingual, where the questions were formulated in a language different from that of the news collection.

Eleven source languages were considered, namely, Bulgarian, Dutch, English, French, German, Indonesian, Italian, Polish, Portuguese, Romanian and Spanish. Note the loss of Finnish, and the introduction of Polish and Romanian with respect to last year. All these languages were also considered as target languages, except for Indonesian, Polish and Romanian. These three languages had no news collection available for the queries. As was done for Indonesian in the previous two campaigns, the English question set was translated into Indonesian (IN), Polish (PL) and

Table 1. Task activated in 2006

		TARGET LANGUAGES (corpus and answers)							
		BG	DE	EN	ES	FR	IT	NL	PT
SOURCE LANGUAGES (questions)	BG								
	DE								
	EN								
	ES								
	FR								
	IN								
	IT								
	NL								
	PL								
	PT								
	RO								
	RO								

Romanian (RO), and the German question set into Romanian (RO). Only the bilingual tasks IN-EN, PL-EN, RO-EN and RO-DE were activated. In the case of IN-EN, PL-EN, and RO-EN, the questions were posed in the respective language (i.e. IN, PL, RO), while the answers were retrieved from the English collection. In the RO-DE case, the question was made in Romanian, whilst the answer was retrieved from the German collection.

As shown in Table 1, 24 tasks were proposed and divided in:

- 7 Monolingual -i.e. Bulgarian (BG), German (DE), Spanish (ES), French (FR), Italian (IT), Dutch (NL), and Portuguese (PT);
- 17 Cross-lingual.

As customary in recent campaigns, a monolingual English (EN) task was not available as it seems to have been already thoroughly investigated in TREC campaigns, even though English was both source and target language in the cross-language tasks.

Although the task was not radically changed with regard to previous campaigns, some new elements were introduced. The most important one was the addition of List questions to the question sets, which implied some major issues. For this first year of QA@CLEF, we were not too strict on the definition of lists, using both questions asking for a specific finite number of answers (that could be called "closed lists") e.g.:

Q: What are the names of the two lovers from Verona separated by family issues in one of Shakespeare's plays?

A: Romeo and Juliet.

and open lists ,where as many correct answers could be returned, e.g.

Q: Name books by Jules Verne.

Table 2. Tasks chosen by at least 1 participant in QA@CLEF campaigns

	MONOLINGUAL	CROSS-LINGUAL
CLEF 2003		
CLEF 2004	6	13
CLEF 2005	8	15
CLEF 2006	7	17

and let organizing groups decide on how to assess the answers to these different kinds of questions.

Other innovations were:

- the input format, where the type of question (F,D,L) was no longer indicated;
- and the result format, where up to a maximum of ten answers per question was allowed, with one to ten text snippets supporting the exact answer.

3 Test Set Preparation

Following the procedure established in previous campaigns, initially each organising group (one for each Target language) was assigned a number of topics taken from the CLEF IR track on which candidates' questions were based. This choice was originally made to reduce the number of duplicates in the multilingual question set. As the number of new topics introduced in 2006 was small, old topics were simply reassigned to different groups. Some groups questioned this methodology, preferring to produce questions with other methods instead of following particular topics. The topics, and hence the questions, were aimed at data collections composed of news articles provided by ELRA/ELDA dating back to 1994/1995; with the exception of Bulgarian, which dated back to 2000 (see Table 3). The choice of a different collection was a matter for long discussion, copyright issues remaining a major obstacle. A step towards a possible solution was nevertheless made by the proposal of the WiQA pilot task, which represents a first attempt to set the QA competitions in their natural context, i.e. the Internet.

Initially, 100 questions were selected in each of the source languages, distributed between Factoid, Definition and List questions.

Factoid questions are fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.

Table 3. Document collections used in CLEF 2006

TARGET LANG..	COLLECTION	PERIOD	SIZE
Bulgarian (BG)	Sega	2002	120 MB (33,356 docs)
	Standart	2002	93 MB (35,839 docs)
Germany (DE)	Frankfurter Rundschau	1994	320 MB (139,715 docs)
	Der Spiegel	1994/1995	63 MB (13,979 docs)
	German SDA	1994	144 MB (71,677 docs)
	German SDA	1995	141 MB (69,438 docs)
English (EN)	Los Angeles Times	1994	425 MB (113,005 docs)
	Glasgow Herald	1995	154 MB (56,472 docs)
Spanish (ES)	EFE	1994	509 MB (215,738 docs)
	EFE	1995	577 MB (238,307 docs)
French (FR)	Le Monde	1994	157 MB (44,013 docs)
	Le Monde	1995	156 MB (47,646 docs)
	French SDA	1994	86 MB (43,178 docs)
	French SDA	1995	88 MB (42,615 docs)
Italian (IT)	La Stampa	1994	193 MB (58,051 docs)
	Italian SDA	1994	85 MB (50,527 docs)
	Italian SDA	1995	85 MB (50,527 docs)
Dutch (NL)	NRC Handelsblad	1994/1995	299 MB (84,121 docs)
	Algemeen Dagblad	1994/1995	241 MB (106,483 docs)
Portuguese (PT)	Público	1994	164 MB (51,751 docs)
	Público	1995	176 MB (55,070 docs)
	Folha de São Paulo	1994	108 MB (51,875 docs)
	Folha de São Paulo	1995	116 MB (52,038 docs)

The following 6 answer types for factoids were considered:

- PERSON (e.g. "Who was Lisa Marie Presley's father?")
- TIME (e.g. "What year did the Second World War finish?")
- LOCATION (e.g. "What is the capital of Japan?")

- ORGANIZATION (e.g. *"What party did Hitler belong to?"*)
- MEASURE (e.g. *"How many monotheistic religions are there in the world?"*)
- OTHER, i.e. everything else that does not fit into the other five categories (e.g. *"What is the most-read Italian daily newspaper?"*)

Definition questions, i.e. questions like "What/Who is X?", were divided into the following categories:

- PERSON -i.e. questions asking for the role, job, and/or important information about someone (e.g. *"Who is Lisa Marie Presley?"*);
- ORGANIZATION -i.e. questions asking for the mission, full name, and/or important information about an organization (e.g. *"What is Amnesty International?"* or *"What is the FDA?"*);
- OBJECT -i.e. questions asking for the description or function of objects (e.g. *"What is a Swiss army knife?"*, *"What is a router?"*);
- OTHER -i.e. question asking for the description of natural phenomena, technologies, legal procedures etc. (e.g. *"What is a tsunami?"*, *"What is DSL?"*, *"What is impeachment?"*).

The last two categories were especially added to reduce the numbers of definition questions which may be answered very easily (such as acronyms concerning organizations, which are usually answered rendering the abbreviation in full, and people's job-description, which are usually found as appositions of proper names in news text).

As mentioned above, questions that require a list of items as answers, were introduced for the first time. (e.g. *Name works by Tolstoy*).

Among these three categories, a number of NIL questions, i.e. questions that do not have any known answer in the target document collection, were distributed. They are important because a good QA system should identify them, instead of returning wrong answers.

Table 4. Test set breakdown according to question type

	F (150)	D (40)	L (10)	T (40)	NIL (20)
BG	145	43	12	26	17
DE	153	37	10	45	20
EN	150	40	10	40	18
ES	148	42	10	40	21
FR	148	42	10	40	20
IT	147	41	12	38	20
NL	147	40	13	30	20
PT	143	47	9	23	18

Three different types of temporal restriction – a temporal specification that provides important information for the retrieval of the correct answer, were associated to a certain number of F, D, L, more specifically:

- restriction by DATE (e.g. "Who was the US president in 1962?"; "Who was Berlusconi in 1994?")
- restriction by PERIOD (e.g. "How many cars were sold in Spain between 1980 and 1995?")
- restriction by EVENT (e.g. "Where did Michael Milken study before enrolling in the university of Pennsylvania?")

The distribution of the questions among these categories is described in Table 4.

Each of the question sets was then translated into English, so that each group could choose additional 100 questions from those proposed by the others and translate them in their own languages. At the end, each source language had 200 questions, which were collected in an XML document. Unlike in the previous campaigns, the questions were not translated in all the languages due to time constraints, and the Gold Standard contained questions in multiple languages only for activated tasks. Since Indonesian, and Romanian did not have a data collection of their own, the English question set was translated, so that the cross-lingual subtasks IN-EN and RO-EN were made available. As not all questions had been previously translated, a translation of the target language question sets into the source languages was needed for cross-language sub-tasks which had at least one registered participant.

4 Participants

The number of participants has constantly grown over the years [see Table 5]. In fact, about ten new groups have joined the competition each year, and in 2006 a total of 30 participants was reached.

Table 5. Number of participating groups

	America	Europe	Asia	Australia	TOTAL	Registered participants	New groups	Veterans	Absent veterans
CLEF 2003	3		-	-	8				
CLEF 2004	1	17	-	-	18 (+125%)				
CLEF 2005	1	22	1	-	24(+33%)	27	9	15	4
CLEF 2006	4	24	2	-	30 (+25%)	36	10	20	4

For the record, the number of groups which registered for the competition but did not actually participate in it was six, while four groups which took part in QA2005 did not show up in 2006. From a geographical perspective, most groups came from

Europe, but in 2006 there was an increase in participants from both Asia and America [see Table 5].

The increase in the number of submitted runs corresponded to that of the participants. Of higher significance is the slight decrease registered in monolingual subtasks to the advantage of bilingual ones. This indicates that QA@CLEF is becoming increasingly cross-lingual, as it was originally set out to be.

Table 6. Number of submitted runs

	Number of submitted runs	Monolingual	Cross-lingual
CLEF 2003	17	6	11
CLEF 2004	48	20	28
CLEF 2005	67	43	24
CLEF 2006	77	42	35

5 Evaluation

The introduction of list questions, the possibility to return multiple answers, and the requirement of supporting the answers with snippets of texts from the relevant documents made the evaluation process more difficult. Moreover, in some languages the large amount of data requiring assessment made it impossible for the judging panels to correct more than one answer per question. Therefore, only the first answers were evaluated in runs that had English and Spanish as a target. In all other cases at least the first three answers were evaluated.

Considering these issues, it was decided to follow the procedure utilised during the previous campaign. The files submitted by the participants in all tasks were manually judged by native speakers. Each language coordination group guaranteed the evaluation of at least one answer per question.

If a group decided to assess more than one answer per question, the answers were assessed in the order they occurred in the submission file and the same number was applied to all questions, and all the runs assessed by the group. The exact answer (i.e. the shortest string of words which is supposed to provide the exact amount of information to answer the question) was assessed as:

- R (Right) if correct;
- W (Wrong) if incorrect;
- X (ineXact) if contained less or more information than that required by the query;
- U (Unsupported) if either the docid was missing or wrong, or the supporting snippet did not contain the exact answer.

Most assessor-groups managed to guarantee a second judgement of all the runs, with a good average inter-assessor agreement. As far as the evaluation measures are concerned, the list questions had to be scored separately, and different groups returned a different number of answers for originally meant Factoid and Definition questions. As a consequence, we decided to provide the following measures:

- accuracy, as the main evaluation score, defined as the average of SCORE(q) over all 200 questions q ;
- the mean reciprocal rank (MRR) over N assessed answers per question. That is, the mean of the reciprocal of the rank of the first correct label over all questions;
- the K1 measure used in earlier QA@CLEF campaigns [4]
- the Confidence Weighted Score (CWS) designed for systems that give only one answer per question. Answers are in a decreasing order of confidence and CWS rewards systems that give correct answers at the top of the ranking [4].

Although some other kinds of measures have been proposed and used in CLEF 2005, such as a more detailed analysis/breakdown of bad answers by the Portuguese group [7], they were not considered this year. Also, issues like providing more accurate description of what X means: too much or too little were only distinguished by the Portuguese assessors, argued for i.a. in Rocha and Santos [6].

6 Results

As far as accuracy is concerned, a general improvement has been noticed, as Figure 1 shows.

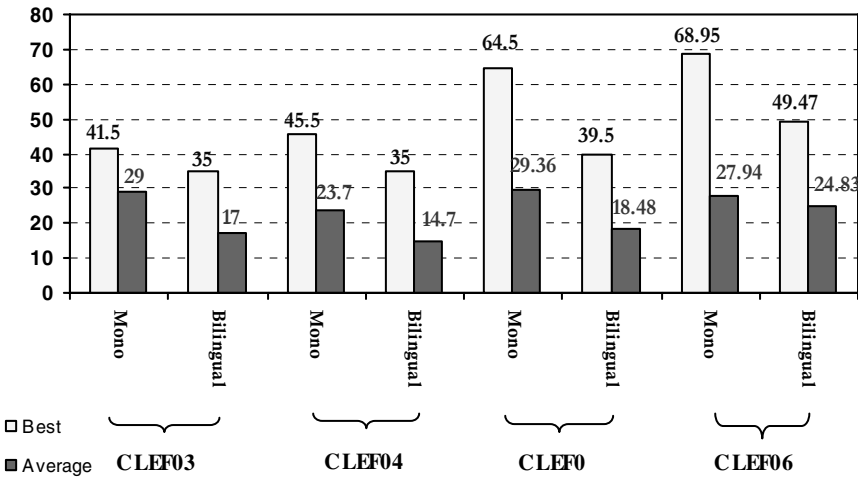


Fig. 1. Best and average scores in CLEF QA campaigns

In detail, Best Accuracy in the monolingual task improved by 6.9%, passing from last year's 64.5% to 68.95%, while Best Accuracy in cross-language tasks passed from 39.5% to 49.47%, recording an increment of 25.2%. As far as average performances are concerned, a slight decrease has been recorded in the monolingual tasks, which went from 29.36% to 27.94%. This probably was due to the number of newcomers which tested their systems for the first time.

As a general remark, best performances has been quite stable, with most languages registering similar or better scores than last campaigns (see Figure 2).

Table 7. Best accuracy scores compared with K1, MRR, and CWS

FILE NAME	OVERALL ACCURACY	K1	MRR	CWS
BEST				
syna061frfr.txt	68.95%	0.2832	0.6895	0.56724
inao061eses.txt	52.63%	0.0716	0.5263	0.43387
ulia061frfr.txt	46.32%	0.0684	0.4632	0.46075
ulia062frfr.txt	45.79%	0.0579	0.4579	0.45546
vein061eses.txt	42.11%	-0.0657	0.4211	0.33582
alia061eses.txt	37.89%	-0.1232	0.3763	0.23630
upv_061eses.txt	36.84%	0.0014	0.3684	0.22530
ulia061enfr.txt	35.26%	-0.1684	0.3526	0.34017

Although also in 2006 campaign self confidence score was not returned by all systems, data about the confidence were plentiful, and allowed to consider the additional evaluation measures, i.e. K1, CWS and MMR. Generally speaking, systems with high accuracy scored accordingly well also with these measures, implying that best systems provide high self confidence, as Table 7 shows.

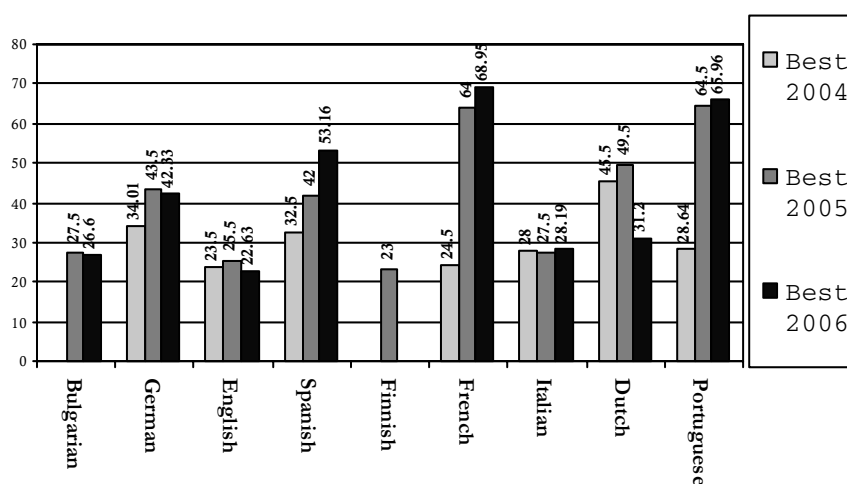


Fig. 2. Best results in 2004, 2005 and 2006

Here below a more detailed analyses of the results in each language follows, giving more specific information on the performances of systems in the single sub-tasks and on the different types of questions, providing the relevant statistics and comments.

6.1 Bulgarian as Target

At CLEF 2006 Bulgarian was addressed as a target language for the second time. This year there was no change in the number of the participants -- again two groups took part in the monolingual evaluation task with Bulgarian as a target language: BTB at Linguistic Modelling Laboratory, Sofia and The Joint Research Centre (JRC), Ispra.

Three runs altogether were submitted – one by the first group and two by the second group with insignificant difference between them. The 2006 results are presented in Table 8 below. First, the correct answers in numbers and percentage are given (Right) per run. Then the wrong (W), inexact (X) and unsupported answers (U) are shown in numbers. Further, the number of the factoids (F), temporally restricted questions (T), definitions (D) and list questions (L) are given. Also, the percentage of the correct answers per each type is registered in Table 8. NIL questions are presented as the number of correctly and wrongly returned answers by the systems with NIL marking. It is obvious that the systems returned NIL answer also when they could not detect a possibly existing answer in the corpus themselves. In our opinion, the present NIL marking might be divided into two labels: NIL = no answer in the corpus is existing and CANNOT = the system itself cannot find an answer. In this way the evaluation would be more realistic. Main reciprocal rank score is provided in the last column of the table.

Table 8a. Results at the Bulgarian as target, monolingual

Run	Right		W #	X #	U #
	#	%			
btb061	50	26.60	132	4	1
jrc061	22	11.70	162	4	0
jrc062	22	11.70	160	6	0

As it can be seen, this year the first system performs better. However, its overall accuracy is slightly worse with respect to the 2005 best accuracy result, achieved then by IRST, Trento. Now it is 26.60 %, while in 2005 it was 27.50 %. However, the BTB 2005 year result was significantly improved. Both systems ‘crashed’ at temporally restricted questions with no single match (see the empty slots in the table). It is a step back from 2005, when both systems had some hits, best of which scored 17.65 %. List questions are also very poorly answered (1 correct answer per run).

The only outperforming results in comparison with the last year are the following: the improvement of the definition type answers (from 42 % to 55.81 %) and the raise of the main reciprocal rank score (from 0.160 to 0.2660).

Table 8b. Results at the Bulgarian as target, monolingual

Run	% F [119]	% T [26]	% D [43]	% L [12]	NIL [16]		r
					right	wrong	
btb061	17.93	-	55.81	0.0833	11	120	0.2660
jrc061	6.90	-	27.91	0.0833	12	155	0.1170
jrc062	6.90	-	27.91	0.0833	13	154	0.1170

The introduction of the snippet support proved out to be a good idea. There was only 1 unsupported answer in all three runs.

The interannotator agreement was very high due to two reasons: first, the number of the answered questions was not very high, and second, there were strict guidelines for the interpretation of the answers, based on our last year experience.

In spite of the somewhat controversial results from the participating systems this year, there is a lot of potential in the task of Bulgarian as a target language in several aspects: investing in the development of the present systems and creating new systems. We hope that Bulgarian will become even more attractive as an EU language.

6.2 Dutch as Target

This year three teams that took part in the CLEF QA track used Dutch as the target language: the University of Amsterdam, the University of Groningen and the University of Roma – 3, with six runs submitted in total: three Dutch monolingual and three crosslingual (English to Dutch). All runs were assessed by two assessors, with the overall inter-assessor agreement 0.96.

Table 9a. Results at the Dutch as target, monolingual

Run	Right		W	X	U
	#	%	#	#	#
Gron061nlnl	58	31.02	115	11	3
Isla061nlnl	40	21.39	141	4	2
Isla062nlnl	41	21.93	139	4	3

For creating the gold standard for Dutch, the assessments were automatically reconciled in favour of more lenient assessments: for example, in case the same answer was assessed as W (incorrect) by one assessor and as X (inexact) by another, the X judgement was included in the gold standard. The results of the evaluation of the six runs are provided in Tables 9 and 10. The columns labelled Right, W, X and U give the results for factoid, definition and temporally restricted questions.

Table 9b. Results at the Dutch as target, monolingual

Run	% F [146]	% T [0]	% D [40]	P@N (lists) [13]	Accuracy NIL [10]	MRR [187]
Gron061nlnl	27.40	0.00	45.00	23.08	0	0.3460
Isla061nlnl	21.23	0.00	22.50	0.00	0.1346	0.2341
Isla062nlnl	21.92	0.00	22.50	0.00	0.1346	0.2357

An interesting thing to notice about this year’s task is that the overall scores of the systems are lower, compared to the last year’s numbers (44% and 50% of correct answers to factoid questions last year).

Table 10a. Results at the Dutch as target, cross-lingual (English to Dutch)

Run	Right		W	X	U
	#	%	#	#	#
Gron061ennl	38	20.32	139	7	3
Roma061ennl	25	13.37	150	6	3
Roma062ennl	25	13.37	149	7	3

This year’s questions were created by annotators who were explicitly instructed to think of “harder” questions, that is, involving paraphrases and some limited general knowledge reasoning. It would be interesting to compare the performance of this year’s systems on last year’s questions to the previous results of the campaign.

Table 10b. Results at the Dutch as target, cross-lingual (English to Dutch)

% F [146]	% T [0]	% D [40]	P@N (lists) [13]	Accuracy NIL [10]	MRR [187]
18.37	0.00	28.21	6.15	0.1481	0.2239
11.56	0.00	20.51	17.95	0.0769	0.1430
11.56	0.00	20.51	15.38	0.0769	0.1529

6.3 English as Target

Creation of Questions. The question for creation of the questions was very similar to last year and is now a well understood procedure. This year it was required to store supporting snippets for the reference answers but this was not difficult and is well worth the trouble. As previously, we were requested to set Temporarily Restricted questions and to distribute these in a prescribed way over the various Factoid question types (PERSON, LOCATION etc). We achieved our quotas but this was extremely difficult to accomplish and we do not feel the time spent is worthwhile as the addition of temporal restrictions more than doubles the time taken to generate the questions.

On the other hand, as the restrictions are frequently synthetic in nature, our knowledge of how to solve these important questions does not necessarily advance from year to year.

Searching for Definition questions (or indeed any questions beyond Factoids) is always very interesting work but the method of evaluation was not clarified this year. So, while the topics we selected do follow the guidelines, we were not required to (or indeed able to) state at generation time exactly what a complete and correct answer should look like. In consequence we can not conclude much from an analysis of the answers returned by systems to such questions.

Summary Statistics for all the Runs. Overall, thirteen cross-lingual runs with English as a target were submitted. The results are shown in. Ten groups participated in seven languages, French, German, Indonesian, Italian, Romanian, Polish and Spanish.¹ There were three groups for French, two for Spanish and one for all the rest

Results Analysis. There were three main types of question this year, Factoids, Definitions and Lists and we consider the results over these types as well as considering the best scores overall. The most indicative measure overall is a simple count of correct answers and this is what we have used. For the 150 Factoids the top four systems were lire062fren (39), lire061fren (33), dlgt061fren (32) and aliv061esen (29). The other results are not greatly different from last year. The top result of 39/150 amounts to 26%. For the 40 definitions, the picture is similar. The top five results are aliv062esen (11), lire061fren (10), aliv061esen (9), lire062fren (9) and dfki061deen (8). For each of the ten list questions, a system could return up to ten candidate answers. Considering both a simple count of correct answers and the P@N score achieved, the top five results by count are uaic061roen (10, 0.11), lire061fren (9, 0.09), irst061iten (8, 0.16), lire062fren (8, 0.08) and dfki061deen (6, 0.2). The ordering for the P@N score differs: dfkienen (6, 0.2) irst061iten (8, 0.16), uaic061roen (10, 0.11), lire061fren (9, 0.09) and lire062fren (8, 0.08).

Assessment Procedure. This approach to assessment was broadly similar to that of last year. However, as the format of the runs had changed, we decided not to use the NIST software but to work with the bare text files instead. It had been intended to double-judge all the questions but unexpectedly and at the last moment this proved not to be possible due to the absence of an assessor. There were 200 questions in all. One assessor judged all answers to questions 1-100 while the other two judged all answers to questions 101-200.

There were considerable practical problems with the assessment of runs this year. Firstly, several runs used invalid run tags. Secondly two of the runs were answering the questions in a completely different order! Thirdly, one question in these two runs was different from the question being answered by the other systems in that position. Fourthly, one run had the fields in the wrong order. Fifthly one run used NULL instead of NIL while another run used nil. Luckily we spotted problems 2 and 3 and

¹ A Polish-English run -'utjp061plen'- was submitted, achieving an overall accuracy of 86%. However, as the system report did not provide enough information to support such a high score, we have decided not to validate this result.

Table 11. Results of English runs

Run	R #	W #	X #	U #	% F [150]	% D [40]	P@N for L [10]	OVERALL ACCURACY %
aliv061esen	38	142	4	6	19.33	22.50	0.0411	20.00
aliv062esen	29	156	3	2	12.00	27.50	0.0200	15.26
aske061esen	10	134	11	34	6.67	0.00	0	5.26
aske061fren	7	135	10	37	3.33	5.00	0.0100	3.68
dfki061deen	34	147	9	0	17.33	20.00	0.2000	17.89
dltg061fren	36	138	14	2	21.33	10.00	0.2000	18.95
irst061iten	24	152	3	11	16.00	0.00	0.1600	12.63
lire061fren	43	138	2	7	22.00	25.00	0.0900	22.63
lire062fren	48	130	2	10	26.00	22.50	0.0800	25.26
uaic061roen	25	150	7	8	15.33	5.00	0.1131	13.16
uaic062roen	18	171	1	0	12.00	0.00	0.0800	9.47
uind061inen	14	159	4	13	9.33	0.00	0	7.37

were able to correct them and indeed all the others but this was extremely time consuming and difficult.

As in all previous years the runs were anonymised by a third party so none of the assessors knew either the origin of a run or the original source language.

This year it had been decided to allow multiple answers to Factoid and Definition questions (up to ten per question). The rationale for this was never quite clear since the whole objective of Question Answering (as against Information Retrieval) is to return only the right answer. Even in cases where there are genuinely several right answers (a rare situation in our carefully designed question sets) a system should still return a correct answer in the first place. For this reason and due to our limited time and resources, we only judged the first answer returned to Factoid and Definition questions. For List questions, all candidate answers were judged, as is normal at TREC.

For the questions double judged, we measured the agreement level. There were 149 differences over thirteen runs of 100 questions. This amounts to 149/1300 i.e. 11% disagreement or 89% agreement. The overall figure for last year was 93%.

Concerning the judgement process itself, Factoids and Lists did not present a problem as we were very familiar with them. On the other hand Definitions were in the same state as last year in that they had been included in the task without a suitable evaluation procedure having been defined. In consequence we used the same approach as last year: If an answer contained information relevant to the question and also contained no irrelevant information, it was judged R if supported, and U

otherwise. If both relevant and irrelevant information was present it was judged X. Finally, if no relevant information was present, the answer was judged W.

Comment and Conclusions. The number of runs judged (12) was the same as last year. However, two source languages were introduced: Indonesian and Romanian. The results themselves were also broadly similar.

Definition questions remained in the same unspecified state as previously. This means that we have not been successful in stretching the boundaries of question answering beyond Factoids which are now very well understood. This is a great pity as the extraction of useful 'definition type' information on a topic is a very useful task for groups to study but it is one which needs to be carefully quantified.

The introduction of snippets was very helpful at question generation time and also invaluable for judging the answers. Snippets are a great step forward for CLEF and are the most significant development for the QA Track this year.

6.4 French as Target

This year (as last year) seven groups took part in evaluation tasks using French as target language: four French groups: Laboratoire d'Informatique d'Avignon (LIA), CEA-List, Université de Nantes (LINA) and Synapse Développement; one Spanish group: Universitat Politècnica de Valencia; one Japanese group; and one American group: LCC.

In total, 15 runs have been returned by the participants: eight monolingual runs (FR-to-FR) and seven bilingual runs (6 EN-to-FR, 1 PT-to-FR).

It appears that the number of participants for the French task is the same that last year but it's the first time there are non-European participants. This shows there is a new major interest for the French as target language.

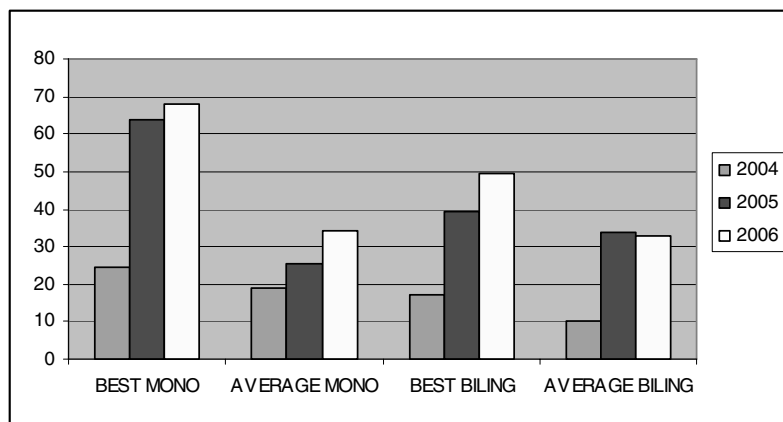


Fig. 3. Best and average scores for systems using French as target in CLEF QA campaigns

Two groups submitted four runs, two other groups submitted two runs and three groups submitted only one run.

Figure 3 shows the best and the average scores for systems using French as target in the last three CLEF QA campaigns.

For both monolingual and bilingual tasks, the best results were obtained by a French group, Synapse Développement. Another French group, LIA, reached the 2nd position for the two tasks.

Table 12 and 13 shows the results of the assessment of each run for each participant and for the two tasks.

The For the monolingual task, the systems returned between 27 and 129 correct answers in 1st rank.

For the bilingual task, the systems returned between 19 and 86 correct answers.

test set was composed of 190 Factual (F), Definition (D) and Temporally restricted (T) questions, and 10 List questions.

The accuracy has been calculated over all the first answers of F, D, T questions and also the Confidence Weighted Score (CWS), the Mean Reciprocal Rank score (MRR) and the K1 measure.

For the List questions, the P@N has been calculated.

For the monolingual task, the best system returned 67.89 % of correct answers (overall accuracy in 1st rank). We can observe this system obtained better results for definition questions (83.33 %) than for Factoid questions (63.51 %).

The LIA' system, which reached the second position in this task, returned 46.32 % of correct answers (overall accuracy in 1st rank). We can also observe the difference between the results for the Factual questions and the results for the Definition questions: 37.84 % of correct answers for the Factual and 76.19 % for the Definition questions.

Table 12a. Assessment of the monolingual and bilingual French

Id Participant	Assessed Answers (#)	Right answers (#)	Wrong answers (#)	ineXact answers (#)	Unanswers (#)
aske061frf	635	27	138	12	12
lcea061frf	589	30	151	6	3
lina061frf	207	56	114	18	2
syna061fr	200	129	50	9	2
ulia061frf	200	88	93	7	2
ulia062frf	200	86	89	9	6
upv061frf	200	60	119	10	1
upv062frf	200	47	124	18	1
aske061en	640	19	157	6	8
lcc061enfr	578	40	125	23	2
syna061e	200	86	97	6	1
syna062en	200	63	120	6	1
ulia061enf	200	66	114	7	3
ulia062enf	200	66	111	9	4
syna061pt	200	94	90	4	2

For the bilingual task, the best system obtained 45.26 % of correct answers as opposed to 34.74 % of correct answers for the LIA' system.

We can remark that the best system for the bilingual task (EN-to-FR) obtained worse results than the second system for the monolingual task.

This year, before the assessment, the French assessors determined some rules to face up to problems encountered the last year.

Concerning Temporally restricted questions for example, to assess an answer as "Correct", the date, the period or the event had to be present in the document returned by the systems.

They decided also to check separately, at the end of the assessment, some questions which seemed difficult to them, to make sure that each answer had received the same "treatment" during the evaluation.

The main problem encountered this year, was related to the assessment of the List questions. This was a new kind of questions this year and participants followed different ways to answer to these questions. Some systems returned a list of answers in a same line; others returned an answer per line. ELDA evaluated these answers according to each run (if a line contained one of correct answers or all the correct answers, these answers had been assessed as "Correct").

The best system obtained 5 correct answers out of 10 List questions in total.

We can observe that the results for the List questions were not very relevant because of not much questions and not much rules.

Table 12b. Assessment of the monolingual and bilingual Frenchv

Participant Id	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	MRR (F, D, T)	CWS (F, D, T)	KI Measure	P@N (L)
aske061frf	14.21	16.89	4.76	0.1974	0.1421	---	0.0900
lcea061frfr	15.79	10.14	35.71	0.1907	0.1578	---	0.1633
lina061frfr	29.47	27.70	35.71	0.2947	0.2551	-0.3777	0.3651
syna061fr	67.89	63.51	83.33	0.6789	0.5568	0.2729	0.5000
ulia061frfr	46.32	37.84	76.19	0.4632	0.4607	0.0684	0.5000
ulia062frfr	45.26	36.49	76.16	0.45016	0.4501	0.0474	0.2000
upv061frfr	31.58	31.08	33.33	0.3158	0.1638	-0.0047	0.3000
upv062frfr	24.74	26.35	19.05	0.2474	0.1088	-0.0931	0.2000
aske061en	10.00	12.16	2.38	0.1445		-0.2797	0.0633
lcc061enfr	21.05	25.00	7.14	0.2623		-0.1816	0.3967
syna061en	45.26	37.16	73.81	0.4526	0.4526	---	0.2000
syna062en	33.16	25.68	59.52	0.3316	0.3315	---	0.1000
ulia061enf	34.74	26.35	64.29	0.3474	0.334	-0.1789	0
ulia062enf	34.74	26.35	64.29	0.3474	0.3474	-0.1789	0.1000
syna061pt	49.47	41.50	76.74	0.4947	0.4947	---	0

In conclusion, this year, a system obtained “excellent” results. Synapse Développement obtained 129 correct answers out of 200 (as opposed to 128 last year).

This system is the best system for the French language. This year, it’s again the dominant system.

In addition, we can observe the same great interest in Question Answering from the European (and now non-European) research community for the tasks using French as target language.

6.5 German as Target

Three research groups submitted runs for evaluation in the track having German as target language: The German Research Center for Artificial Intelligence (DFKI), FernUniversität Hagen (FUHA) and The Institute for Natural Language Processing in Stuttgart (IMS).

Table 13. Best and Aggregated Mono; Best and Aggregated Cross

Year	Best Mono	Aggregated Mono	Best Cross	Aggregated Cross
2006	42.33	64.02	32.98	33.86
2005	43.5	58.5	23	28
2004	34.01	43.65	0	0

All of them provided system runs for the monolingual scenario and just one group (DFKI) submitted runs for the cross-language English-German scenario.

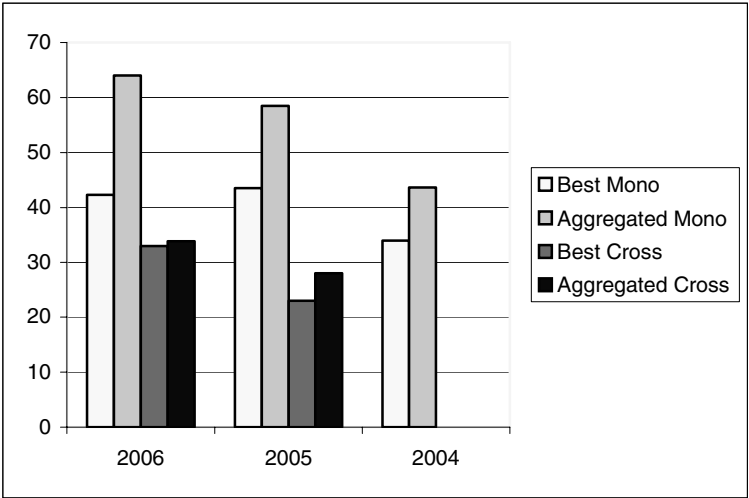


Fig. 4. Results evolution

Table 14. Performance of evaluated systems

Run ID	# Right			# inExact			# Unsupported			Accuracy	MRR
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd		
<i>dfki061dede_M</i>	80	8	7	6	6	1	8	4	1	42.32	45.67
<i>dfki062dede_M</i>	63	15	3	4	5	3	8	0	2	33.33	37.83
<i>fuha061dede_M</i>	61	0	0	0	0	0	4	0	0	32.27	32.27
<i>fuha062dede_M</i>	64	0	0	1	0	0	4	0	0	33.86	33.86
<i>ims061dede_M</i>	25	2	3	0	1	0	8	2	0	13.22	14.28
<i>ims062dede_M</i>	23	3	2	0	1	0	8	2	0	12.16	13.31
<i>dfki061ende_C</i>	62	5	7	3	4	2	6	3	0	32.8	35.36
<i>dfki062ende_C</i>	50	10	2	5	4	2	3	2	1	26.45	29.45

Two assessors with different profiles conducted the evaluation: a native German speaker with little knowledge of QA systems and a researcher with advanced knowledge of QA systems and a good command of German. Compared to the previous editions of the evaluation forum, this year an increase in the performance of an aggregated virtual system for both monolingual and cross-language tasks was registered, as well as for the cross-language best system's result (Figure 4, Table 13). Given the increased complexity of the task (no question type provided, supporting snippets required) and of questions (factoid, definition and list), the stability of the best monolingual results can be considered also a gain in terms of performance.

Table 15a. System Performance: details

Run ID	Right		W	X	U	% F	% T	% D
	#	%	#	#	#	[152]	[44]	[37]
<i>dfki061dede_M</i>	80	42.32	95	6	8	38.81	29.54	56.75
<i>dfki062dede_M</i>	63	33.33	114	4	8	30.92	22.72	43.24
<i>fuha061dede_M</i>	61	32.27	124	0	4	30.26	15.9	40.54
<i>fuha062dede_M</i>	64	33.86	120	1	4	31.57	18.18	43.24
<i>ims061dede_M</i>	25	13.22	156	0	8	13.81	9	10.81
<i>ims062dede_M</i>	23	12.16	158	0	8	12.5	6.81	10.81
<i>dfki061ende_C</i>	62	32.8	117	3	6	28.94	22.72	48.64
<i>dfki062ende_C</i>	50	26.45	130	5	3	22.36	20.45	43.24

Concerning factoid questions, their increased complexity is reflected in systems being required to use some sort of lexical inference in order to track down the relevant contexts of the right answer, contexts that may be scattered across several adjoining sentences.

Table 15b. System Performance: Details

Run ID	% D	P@N L	NIL [20]			CWS	K1
	[37]	[9]	F	P	R		
<i>dfki061dede_M</i>	63.64	25.93	0.35	0.28	0.45	0	0
<i>dfki062dede_M</i>	48.48	33.33	0.32	0.27	0.4	0	0
<i>fuha061dede_M</i>	36.36	11.11	0.23	0.13	0.95	0.3	0.18
<i>fuha062dede_M</i>	39.39	11.11	0.24	0.14	0.95	0.32	0.19
<i>ims061dede_M</i>	9.09	25.42	0.2	0.12	0.55	0.07	-0.33
<i>ims062dede_M</i>	9.09	26.43	0.19	0.12	0.5	0.06	-0.33
<i>dfki061ende_C</i>	56.25	10	0.31	0.21	0.6	0	0
<i>dfki062ende_C</i>	50	10	0.33	0.22	0.65	0	0

Except for FUHA, the other two groups provided more than one possible answer per question, of which only the first three were manually evaluated. In order to come up with a measure of performance for systems providing several answers per question, Mean Reciprocal Rank (MRR) over right answers has been considered for this purpose.

Table 14 resumes the distribution of the right, inexact and unsupported answers over the first three ranked positions as delivered by the systems, as well as the accuracy and MRR for each of the runs.

Table 16. Inter-Assessor Agreement/Disagreement (breakdown)

Run ID	# Questions	# Answers	# Q-Disagreements				# A-Disagreements			
			Total	F	D	L	Total	X	U	W/ R
<i>dfki061dede_M</i>	198	437	35	28	7	0	44	20	16	8
<i>dfki062dede_M</i>	198	476	28	19	6	3	40	13	19	8
<i>fuha061dede_M</i>	198	198	12	8	4	0	11	3	2	6
<i>fuha062dede_M</i>	198	198	13	8	5	0	12	4	2	6
<i>ims061dede_M</i>	198	432	15	13	0	2	30	13	9	8
<i>ims062dede_M</i>	198	436	17	15	0	2	28	5	14	9
<i>dfki061ende_C</i>	198	405	26	20	5	1	33	12	16	5
<i>dfki062ende_C</i>	198	402	27	21	6	0	35	21	10	4

Two things can be concluded from the answer distribution of Table 14: first, there are a fair number of inexact and unsupported answers that show performance could be improved with a better answer extraction; second, the fair number of right answers among the second and third ranked positions indicate that there is still place for improvements with a more focused answer selection.

The details of systems' results can be seen in Table 15, in which the performance measures has been computed only for the first ranked answers to each question, except for the list questions. As an overall remark on the results, the majority of the evaluated system runs registered a better performance on the definition than on the factoid and temporal questions.

Table 16 describes the inter-rater disagreement on the assessment of answers in terms of question and answer disagreement. Question disagreement reflects the number of questions on which the assessors delivered different judgments and answer disagreement is a figure of the total number of answers disagreed on. Along the total figures for both types of disagreement, a breakdown at the question type level (Factoid, Definition, List) and at the assessment value level (ineXact, Unsupported, Wrong/Right) is listed.

The answer disagreements of type Wrong/Right are trivial errors during the assessment process when a right answers was considered wrong by mistake and the other way around, while those of type X or U reflect different judgments whereby an assessor considered an answer inexact or unsupported while the other marked it as right or wrong.

6.6 Italian as Target

Two groups participated in the Italian monolingual task, ITC-irst and the Universidad Politécnica de Valencia (UPV); while one group, the Università La Sapienza di Roma, participated in the cross-language EN-IT task. In total, five runs were submitted.

For the first time a cross-language task with Italian as target was chosen to test a participating system.

The best performance in the monolingual task was obtained by the UPV, which achieved an accuracy of 28.19%. Almost the same result was recorded last year (see Figure 5). The average accuracy in the monolingual task was 26.41%, which is an improvement of more than 2% with respect to last year's results.

The accuracy in the bilingual task was 17.02%, achieved by both submitted runs.

During the years the overall accuracy has steadily decreased starting from a 25.17% in the 2004, we reached a 24.08% in the 2005 and 22.06% this year. This could be partly due to newcomers – who usually get lower scores – and first experiments with bilingual tasks.

From the results shown in Table 17, it can be seen that the Universidad Politécnica de Valencia (UPV) submitted two runs in the monolingual task and achieved the best overall performance. The accuracy over Definition and Factoid questions ranged from 26.83% to 29.27%. ITC-irst submitted one run, and achieved much better accuracy over Factoid questions (25.00%) than over Definition questions (17.07%).

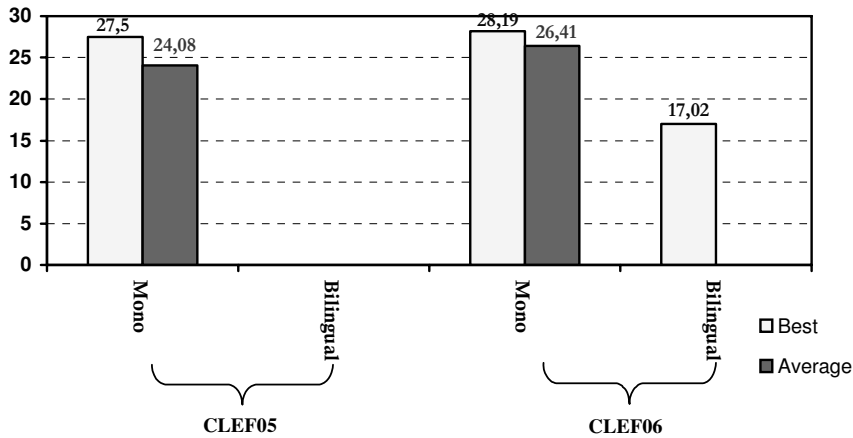


Fig. 5. Best and Average performance in the Monolingual and Bilingual tasks

As previously mentioned, the Università La Sapienza di Roma submitted two runs in the cross-language EN-IT tasks, performing much better in the Definition questions (24.39%) than in the Factoid questions (15.28%).

Table 17. Results of the monolingual and bilingual Italian runs

Run Name	Right answers (#)	Wrong answers (#)	Incorrect answers (#)	Unsupported answers (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	P@N for L	Confidence Weighted Score
irst06itit	43	121	10	13	22.87	25.0	17.07	0.1528	0.19602
upv_061itit	53	124	6	5	28.19	28.47	26.83	0.0833	0.12330
upv_062itit	53	127	4	3	28.19	27.78	29.27	0.1667	0.13209
Roma061enit	32	141	4	11	17.02	15.28	24.39	0.1000	0.08433
Roma062enit	32	141	4	11	17.02	15.28	24.39	0.1500	0.08433

As far as List questions are concerned, all participating systems performed rather poorly, with a P@N ranging from 0.08 to 0.17. This implies that a more in-depth research on these questions and the measures for their evaluation is still needed.

Table 18. Temporally Restricted Questions: Right, Unsupported and Wrong Answer and Accuracy

	R	U	W	Accuracy %
irst06itit	6	6	26	15.79
Roma061enit	2	4	32	5.26
Roma062enit	2	4	32	5.26
upv_061itit	8	0	30	21.05
upv_062itit	9	0	29	23.68

Temporally restricted questions represented a challenge for the systems, which generally achieved a lower than average accuracy in this sub-category. The Universidad Politécnica de Valencia achieved the best performance of 23.68% (see Table 18). The evaluation process did not presented particular problems, although it was more demanding than usual because of the necessity to check the supporting text snippet. All runs were anyway assessed by two judges. The inter-assessor agreement was averagely 90,14 %, most disagreement being between U and X. A couple of cases of disagreement between R and W were due just to trivial mistakes.

6.7 Portuguese as Target

This year five research groups took part in tasks with Portuguese as target language, submitting ten runs: seven in the monolingual task, two with English as source, and one with Spanish. Two new groups joined for Portuguese: University of Porto, and Brazilian NILC, while LCC participated with an English-Portuguese run only. Universidade de Évora did not participate this year.

Table 19 presents the overall results concerning all questions. We present values both taking into account only the first answer to each question, and – for the only system where this makes any difference – all answers, assessing as right (or partially right) if any answer, irrespective of position, was right (or partially right). We have also distinguished inexact answers (X) between too little and too much information, respectively coded as X- and X+. There were only 18 NIL questions in the Portuguese collection.

Just like last year, Priberam achieved the best results by a clear margin. Also, their Spanish-Portuguese run, *prib061espt*, despite using a different (closely related) language as source, managed to achieve the second best result.

On the other hand, overall results for both Priberam and Esfinge show but a small improvement compared to 2005

It remains to be seen whether this year's questions displayed a higher difficulty or whether the systems themselves were subject to few changes.

We also provide in Table 20 the overall accuracy considering (and evaluating) independently all different answers provided by the systems.

Table 19. Results of the runs with Portuguese as target: first answers only, and all answers (marked with *)

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy	
									Pre cision	Rec all
esfg061ptpt	49	139	7	2	3	24.5	22.88	29.79	15.53	88.89
esfg062ptpt	45	142	6	6	1	22.5	20.26	29.79	14.95	88.89
nilc061ptpt	0	189	1	8	2	0.0	0.00	0.00	-	-
nilc062ptpt	3	190	0	5	2	1.5	1.96	0.00	8.57	16.67
prib061ptpt	134	58	6	1	1	67.0	65.36	72.34	43.33	72.22
uporto061ptpt	23	177	0	0	0	11.5	9.80	17.02	8.29	77.78
uporto062ptpt	26	169	2	3	0	13.0	11.76	17.02	7.64	66.67
esfg061enpt	27	164	3	2	2	13.5	11.76	19.15	11.35	88.89
lcc_061enpt	18	166	2	10	4	9.0	9.15	8.51	21.43	16.67
lcc_061enpt*	61	112	3	18	7	30.5	36.1	12.5	-	-
prib061espt	67	124	2	2	5	33.5	26.97	51.06	17.65	33.33

Table 20. Results of the runs with Portuguese as target: all answers

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy %)
esfg061ptpt	49	143	11	2	3	23.56
esfg062ptpt	45	146	7	6	1	21.95
nilc061ptpt	0	189	1	8	2	0.00
nilc062ptpt	3	190	0	5	2	1.50
prib061ptpt	134	58	6	1	1	67.00
uporto061ptpt	36	178	0	0	0	16.82
uporto062ptpt	42	172	3	6	0	18.83
esfg061enpt	27	168	3	2	2	13.36
lcc_061enpt	141	1209	11	49	50	9.66
prib061espt	67	124	2	2	5	33.50

Table 21. Results of the assessment of the monolingual Portuguese runs: first answers only, except for lists, for which (for this table) one correct member of the list means the answer is to be considered correct.

Run	correct answers											
	Definition (47)				Factoid (t.r.q. ; list) (153 (27; 9))						Total	
	Object 7	Organisation 8	Other: 24	Person 9	Location 25 (2:0)	Measure21 (2:0)	Organisation23 (6:3)	Other 30 (6:2)	Person 34 (11:3)	Time 19 (0:1)	# 200 (27:9)	%
esfg061pt pt	3	3	5	3	10	4 (1:0)	2 (1:0)	6 (0:0)	11 (2:1)	3(0:0)	49 (4:1)	24.5
esfg062pt pt	3	4	5	2	9	4 (1:0)	1 (1:0)	6 (1:0)	8 (2:1)	3 (0:0)	45 (4:1)	22.5
nilc061pt pt	0	0	0	0	0	0	0	0	0	0	0	0
nilc062pt pt	0	0	0	0	1	1 (1:0)	0	0 (1:0)	0	1(0:0)	3 (1:0)	1.5
prib061pt pt	5	6	14	8	17 (1:0)	3 (1:0)	17 (3:3)	17 (1:1)	20 (4:3)	16 (0:1)	134 (10:8)	67.0
uporto061ptpt	1	0	2	5	3	3 (1:0)	1 (1:0)	2 (0:0)	5 (2:1)	1 (0:0)	23 (4:1)	1.5
uporto062ptpt	1	0	2	5	4	3 (1:0)	1 (1:0)	3 (0:0)	4 (2:1)	3(0:0)	26 (4:1)	3.0
combination	5	7	19	9	19 (1:0)	4 (1:0)	17 (3:3)	19 (1:1)	23 (5:3)	17 (0:1)	9 (11:8)	74.5
esfg061enpt	1	2	2	4	7	3 (1:0)	1 (1:1)	4 (0:0)	2 (2:1)	1 (0:0)	27 (4:2)	15.4
lcc_061enpt	1	1	2	0	2	1	2 (1:0)	1 (0:1)	2 (1:0)	6 (0:0)	18 (2:1)	9.0
prib061espt	2	4	10	8	7	9 (1:0)	7 (0:0)	6 (1:0)	7 (2:0)	7 (0:0)	67 (3:0)	35.6

Lcc's performance slightly increases, from 9.0 to 9.66, but if we are interested in correct answers anywhere, then Lcc* got the second best accuracy, 30.5, which might suggest that they might significantly improve their score by reranking mechanisms.

Table 21 shows the results for each answer type. In parentheses we display the subset of temporally-restricted questions, and we add the list questions, in order to provide the full picture.

Table 22. Size of answers and justifying snippets, in words

Run name	Answers (#)	Non-NIL Answers (#)	Average answer size	Average answer size (R only)	Average snippet size	Average snippet size (R only)
esfg061ptpt	208	105	3.4	3.2	108.8	108.5
esfg062ptpt	204	98	3.8	3.5	109.1	105.5
nilc061ptpt	200	200	5.7	-	5.7	-
nilc062ptpt	200	165	4.9	-	4.4	-
prib061ptpt	200	170	3.7	3.8	31.5	30.3
uporto061ptpt	210	29	3.1	3.2	39.7	32.7
uporto062ptpt	216	59	3.0	2.8	43.1	33.7
esfg061enpt	202	61	3.5	3.5	95.3	106.1
lcc_061enpt	1463	1449	5.2	4.1	35.2	34.6
prib061espt	200	166	3.5	4.3	31.3	29.1

A virtual run, called combination, was included in Table 21 and computed as follows: if any of the participating systems found a right answer, it is considered right in the combination run. Ideally, this combination run measures the potential achievement of cooperation among all participants. However, for Portuguese this combination does not significantly outperform the best performance: Priberam alone corresponds to 89.9% of the combination run.

Table 23. Accuracy of temporally restricted questions (all answers considered), compared to non-temporally restricted ones, and to overall accuracy

Run name	Questions with at least one correct answer (#)	Accuracy for T.R.Q. (%)	Accuracy for non-T.R.Q (%)	Total accuracy (%)
esfg061ptpt	3	11.11	25.41	23.56
esfg062ptpt	3	11.11	23.60	21.95
nilc061ptpt	0	0.00	0.00	0.00
nilc062ptpt	1	2.86	1.16	1.50
prib061ptpt	10	37.04	71.68	67.00
uporto061ptpt	4	14.81	17.11	16.82
uporto062ptpt	4	14.81	19.39	18.83
esfg061enpt	3	11.11	13.71	13.37
lcc_061enpt	7	2-86	11.03	9.66
prib061espt	3	11.11	36.99	33.50

Table 24. Results for List Questions

Question	Known answers	Esfe061ppt	esfe062ppt	nilc061ppt	nilc062ppt	pnb061ppt	upor061ppt	upor062ppt	esfe061empt	lc061empt	esfe061espt
² 05	3	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/10	0/1
³ 99	3	0/1	0/1	0/1	0/1	3/3	0/1	0/1	0/1	3/9	0/1
⁴ 00	3	0.5/3	0.5/3	0/1	0/1	3/3	0/1	0/1	0.5/3	0/8	3/3
⁷ 59	3	0/1	0/1	0.5/1	0.5/1	1/1	0/1	0/1	0/1	0.5/10	0/1
⁷ 70	3	0.5/1	0.5/1	0/1	0/1	1/1	0/1	0/1	0/1	2/10	0/1
⁷ 84	5	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1	1/9	0/1
⁷ 85	3	0/1	0/1	0/1	0/1	0.5/1	0/1	0/1	0/1	0/10	0/1
⁷ 86	3	0/1	0/1	0.5/1	0.5/1	1/1	0/1	0/1	0/1	2/10	0/1
⁷ 95	5	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1	2/7	0/1
^s core		0.030	0.030	0.037	0.037	0.396	0	0	0.019	0.113	0.011

We have also analysed the size in words of both answers and justification snippets, as displayed in Table 22. (Computations were made excluding NIL answers.) Interestingly, Priberam provided the shortest justifications.

In Table 23, we compare the accuracy of the systems for the 22 temporally restricted questions in the Portuguese question set with their scores for non-temporally restricted ones and their overall performance.

Finally, a total of nine questions were defined by the organization as requiring a list as proper answer. The fact that the systems had to find out whether multiple or single answers were expected was a new feature this year and was not conveniently handled by most systems.

In fact, two systems (Priberam and NILC) completely ignored this and provided a single answer to every question, while two other systems, although attempting to deal with list questions, seemed to fail in appropriately identifying them: RAPOSA (UPorto) provided multiple answers only to non-list questions, and Esfinge produced 11 answers for the nine questions.

In fact, only LCC presented multiple answers systematically, yielding an average of 7.32 answers per question, while no other group exceeded 1.1. For the case of closed lists (where "one" answer might bring all answers, such as "Lituânia, Estónia e Letónia"), we still counted the number of answers individually (3).

We believe further study should be devoted to the list questions for the next years, since a distinction between closed lists and open lists, although acknowledged, was not properly taken into consideration. We have thus chosen to handle all these questions alike, assigning them the following accuracy score: number of correct answers (where X counted as 1/2), presented left of the slash, divided by the sum of the number of existing answers in the collections and the number of wrong distinct answers provided by the system, right of the slash. The results are displayed in Table 24.

6.8 Spanish as Target

The participation at the Spanish as Target subtask is still growing. Nine groups, two more than the last year, submitted 17 runs: 12 monolingual, 3 from English, 1 from French and 1 from Portuguese. Table 25 and Table 26 show the summary of systems results for monolingual and cross-lingual respectively. The number of Right, Wrong (W), Inexact (X) and Unsupported (U) answers

Table 25a. Results at the Spanish as target, monolingual

Run	Right		W #	X #	U #	% F [108]	% T [40]	% D [42]	% L [10]
	#	%							
pribe061	105	52,50	86	4	5	55,56	30,00	69,05	40,00
inao061	102	51,00	86	3	9	47,22	35,00	83,33	20,00
vein061	80	40,00	112	3	5	32,41	25,00	83,33	-
alia061	72	36,00	105	15	8	38,89	22,50	50,00	-
upv_061	70	35,00	119	5	6	37,04	25,00	47,62	-
upv_062	57	28,50	123	6	14	27,78	25,00	40,48	-
aliv061	56	28,00	123	8	13	29,63	22,50	35,71	-
aliv062	56	28,00	132	6	6	26,85	25,00	40,48	-
mira062	41	20,50	148	4	7	21,30	17,50	23,81	10,00
sinaiBruja06	39	19,50	146	6	9	16,67	17,50	33,33	-
mira061	37	18,50	154	3	6	21,30	15,00	16,67	10,00
aske061	27	13,50	143	1	29	15,74	12,50	9,52	10,00

Table 25b. Results at the Spanish as target, monolingual

Run	NIL [20]			r	% Answer Extraction
	F	P	R		
pribe061	0,44	0,34	0,60	-	84,68
inao061	0,46	0,38	0,60	0,216	86,44
vein061	0,34	0,21	0,80	0,133	86,02
alia061	0,34	0,22	0,75	0,322	69,23
upv_061	0,43	0,33	0,65	0,194	70,71
upv_062	0,41	0,32	0,60	0,163	66,28
aliv061	0,34	0,33	0,35	0,190	65,12
aliv062	0,33	0,26	0,45	0,153	72,73
mira062	0,35	0,35	0,35	0,145	43,62
sinaiBruja06	0,23	0,13	0,90	-0,119	79,59
mira061	0,34	0,26	0,50	0,136	51,39
aske061	0,08	0,20	0,05	0,199	62,79

Tables show also the accuracy (in percentage) of factoids (F), factoids with temporal restriction (T), definitions (D) and list questions (L). Best values are marked in bold face. Best performing systems have improved their performance (as seen in Figure 5), mainly with respect to factoids

However, performance when the question has a temporal restriction didn't vary significantly. Last year, the answering of definitions with respect to persons and organizations was almost solved. In spite of the fact that this year the set of definition questions was more realistic systems have improved slightly their performance.

List questions have been introduced this year so they deserve some attention regarding their evaluation. We have differentiated two types of list questions: conjunctive and disjunctive (as presented in [4]).

Conjunctive list questions are asking for a set of items and they are *Right* if all the items are present in the answer. For example, "*Nombre los tres Beatles que siguen vivos*" (Name the three Beatles alive). Disjunctive list questions are asking for an undetermined number of items. For example, "*Nombre luchadores de Sumo*" (Name Sumo fighters). Only the first answer of each system has been evaluated in both cases.

Table 26a. Results at the Spanish as target, Cross-lingual

Run	Right		W	X	U
	#	%	#	#	#
pribe061ptes	72	36,00	123	3	2
alia061enes	41	20,50	134	9	16
lcc_061enes	38	19,00	141	14	7
aske061fres	23	11,50	162	-	15
aske061enes	12	6,00	178	-	10

Table 26b. Results at the Spanish as target, Cross-lingual

Run	% F	% T	% D	% L	NIL [20]			r	% Answer Extraction
	[108]	[40]	[42]	[10]	F	P	R		
pribe061ptes	39,81	27,50	38,10	20	0,29	0,29	0,30	-	78,26
alia061enes	17,59	12,50	40,48	-	0,31	0,19	0,80	0,142	65,08
lcc_061enes	20,37	25,00	14,29	-	0,35	0,35	0,35	0,067	55,07
aske061fres	13,89	10,00	7,14	10	0,08	0,17	0,05	0,302	53,49
aske061enes	6,48	2,50	7,14	10	0,10	1,00	0,05	0,091	40,00

Regarding the NIL questions, Table 25 and 26 show the harmonic mean (F) of precision (P) and recall (R). The best performing systems have increased again their performance (see Table 27) in NIL questions. The correlation efficient r between the self-score and the correctness of the answers has been increased in the majority of systems, although results are not good enough yet.

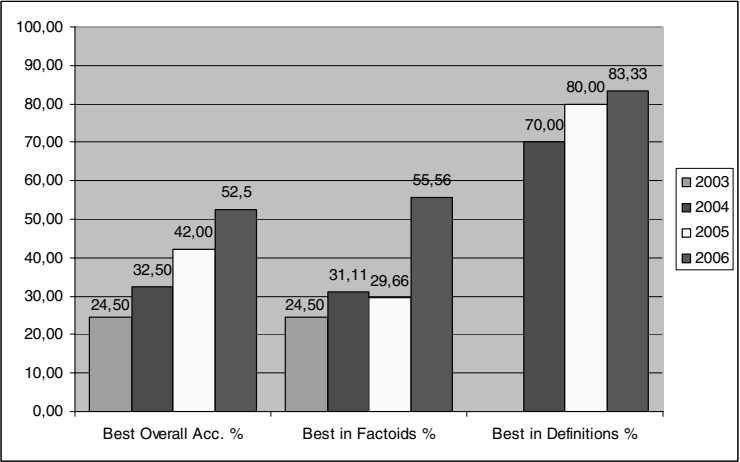


Fig. 6. Evolution of best performing systems 2003-2006

This year a supporting text snippet was requested. For this reason, we have evaluated the systems capability to extract the answer when the snippet contains it.

The last column of Tables 25 and 26 shows the percentage of cases where the correct answer was correctly extracted. This information is very useful to diagnose if the lack of performance is due to the passage retrieval or to the answer extraction.

Regarding Cross-Lingual runs, it is worth to mention that Priberam has achieved in the Portuguese to Spanish task a result comparable to the monolingual runs.

Table 27. Evolution of best results in NIL questions

Year	F-measure
2003	0,25
2004	0,30
2005	0,38
2006	0,46

All the answers have been assessed anonymously considering all systems' answers simultaneously question by question. The inter-annotator agreement was evaluated over 985 answers assessed by the two judges. Only a 2.5% of the judgements were different and the resulting kappa value was 0.93.

7 Conclusions

The QA track at CLEF 2006 has once again demonstrated the interest for Question Answering in languages other than English. In fact, both the number of participants and runs submitted has grown, following the positive trends of the previous campaign.

The balance between tradition and innovations –i.e the introduction of list questions and supporting text snippets- has proved to be a good solution, which allows both new-comers and veterans to test their systems against adequately challenging tasks and, at the same time, to make a comparison with previous exercises. Generally speaking, the results recorded an improvement in performance, with best accuracy significantly higher than in previous campaigns both in monolingual and bilingual tasks.

As far as the organisation of the campaign is concerned, the introduction of new elements such as list questions and supporting snippets has implied a significant increase of work both in the question collection and in the evaluation phase, which was particularly demanding for language groups which had a great number of participants. A better distribution of the workload and solutions to speed up the evaluation process, also with automatic assessment of part of the submissions will be essential in next campaigns.

A future perspective of QA is certainly outlined by the two pilot tasks offered in 2006-i.e. AVE and WiQa-, the latter in particular representing a significant step toward a more realistic scenario, where queries are carried out on the Web. For these reasons, a quick integration of these experiments into the main task is hoped for.

Acknowledgements

The authors would like to thank Donna Harman for her valuable feedback and advice, and Diana Santos for her precious contribution in the organization of the campaign and the revision of this paper.

Paulo Rocha is thankful to the many useful comments and overall discussion with Diana Santos for the Portuguese part. He was also supported by the Portuguese Fundação para a Ciência e Tecnologia within the Linguatca project, through grant POSI/PLP/43931/2001, co-financed by POSI.

Bogdan Sacaleanu was supported by the German Federal Ministry of Education and Research (BMBF) through the projects HyLaP and COLLATE II.

Anselmo Peñas has been partially supported by the Spanish Ministry of Science and Technology within the project R2D2-SyEMBRA (TIC-2003-07158-C04-02).

References

1. QA@CLEF 2006 Organizing Committee. Guidelines (2006), <http://clef-qa.itc.it/guidelines.html>
2. WiQA Website, <http://ilps.science.uva.nl/WiQA/>
3. AVE Website, <http://nlp.uned.es/QA/AVE/>

4. Herrera, J., Peñas, A., Verdejo, F.: Question answering pilot task at CLEF 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 581–590. Springer, Heidelberg (2005)
5. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question Answering Track. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 371–391. Springer, Heidelberg (2005)
6. Rocha, P., Santos, D.: Abrindo a porta á participa internacional em avaliação de RI do português. In: Santos, D. (ed.) Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, IST Press, Lisbon (in Press, 2006)
7. Santos, D., Rocha, P.: The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 821–832. Springer, Heidelberg (2005)
8. Spark Jones, K.: Is question answering a rational task? In: Bernardi, R., Moortgat, M. (eds.) Questions and Answers: Theoretical and Applied Perspectives. Second CoLogNETElsNET Symposium, pp. 24–35. Amsterdam (2003)
9. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 Multilingual Question Answering Track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 21–23. Springer, Heidelberg (2006)
10. Voorhees, E. M.: Overview of the TREC 2002 Question Answering Track. In: Voorhees, E.M., Buckland, L.P. (eds) Proceedings of the Eleventh Text Retrieval Conference (TREC 2002 NIST Special Publication 500-251, pp. 115–123. Washington DC (2002)

Overview of the Answer Validation Exercise 2006

Anselmo Peñas, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo,alvarory,vsama,felisa}@lsi.uned.es

Abstract. The first Answer Validation Exercise (AVE) has been launched at the Cross Language Evaluation Forum 2006. This task is aimed at developing systems able to decide whether the answer of a Question Answering system is correct or not. The exercise is described here together with the evaluation methodology and the systems results. The starting point for the AVE 2006 was the reformulation of Answer Validation as a Recognizing Textual Entailment problem, under the assumption that the hypothesis can be automatically generated instantiating hypothesis patterns with the QA systems' answers. 11 groups have participated with 38 runs in 7 different languages. Systems that reported the use of Logic have obtained the best results in their respective subtasks.

1 Introduction

The first Answer Validation Exercise (AVE 2006) was activated to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by QA systems. This automatic Answer Validation is expected to be useful for improving QA systems performance, helping humans in the assessment of QA systems output, improving systems confidence self-score, and developing better criteria for collaborative systems.

Systems must emulate human assessment of QA responses and decide whether an answer is correct or not according to a given snippet. The first AVE has been reformulated as a Textual Entailment problem [1][2] where the hypotheses have been built semi-automatically turning the questions plus the answers into an affirmative form.

Participant systems must return a value YES or NO for each pair text-hypothesis to indicate if the text entails the hypothesis or not (i.e. the answer is correct according to the text). Systems results were evaluated against the QA human assessments.

Participant systems received a set of text-hypothesis pairs that were built from the QA main track responses of the CLEF 2006. The methodology for building these collections was described in [6]. The development collections were built from the QA assessments of last campaigns [3][4][5][7] in English and Spanish. A subtask per language was activated: English, Spanish, French, German, Dutch, Italian, Portuguese and Bulgarian.

The training collections together with the 8 testing collections (one per language) resulting from the first AVE 2006 are available at <http://nlp.uned.es/QA/ave> for researchers registered at CLEF.

Section 2 describe the test collections. Section 3 motivates the evaluation measures. Section 4 presents the results in each language and Section 5 presents some conclusions and future work.

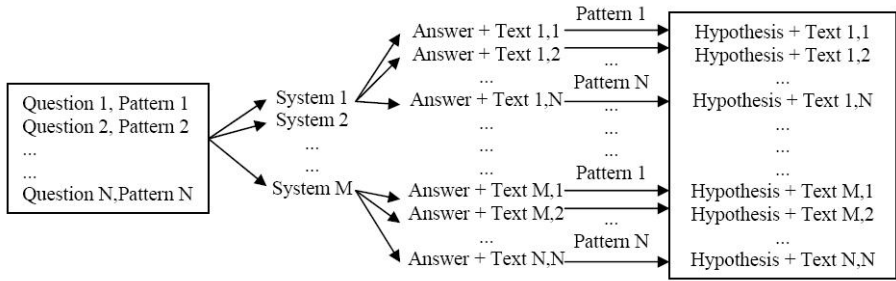


Fig. 1. Text-hypothesis pairs for the Answer Validation Exercise from the pool of answers of the main QA Track

2 Test Collections

As a difference with the previous campaigns of the QA track, a text snippet was requested to support the correctness of the answers. The QA assessments were done considering the given snippet, so the direct relation between QA assessments and RTE judges was preserved: Pairs corresponding to answers judged as Correct have an entailment value equal to YES; pairs corresponding to answers judged as Wrong or Unsupported have an entailment value equal to NO; and pairs corresponding to answers judged as Inexact have an entailment value equal to UNKNOWN and are ignored for evaluation purposes. Pairs coming from answers not evaluated at the QA Track are also tagged as UNKNOWN and they are also ignored in the evaluation.

Figure 1 resumes the process followed in each language to build the test collections. Starting with the 200 questions, a hypothesis pattern was created for each one, and instantiated with all the answers of all systems for the corresponding question (see [6] for more details). The pairs were completed with the text snippet given by the system to support the answer.

Table 1 shows the number of pairs for each language obtained as the result of the processing. This pairs conform the test collections for each language and a benchmark for future evaluations.

The percentages of UNKNOWN pairs are similar in all languages except for the pairs in English and Portuguese, in which up to 5 runs were not finally assessed in the QA task and therefore, the corresponding pairs could not be used to evaluate the systems.

Table 1. YES, NO and UNKNOWN pairs in the testing collections of AVE 2006

	German	English	Spanish	French	Italian	Dutch	Portuguese
YES pairs	344(24%)	198(9.5%)	671(28%)	705(22%)	187(16%)	81(10%)	188(14%)
NO pairs	1064(74%)	1048(50%)	1615(68%)	2359(72%)	901(79%)	696(86%)	604(46%)
UNKNOWN	35(3%)	842(40.5%)	83(4%)	202(6%)	52(5%)	30(4%)	532(40%)
Total	1443	2088	2369	3266	1140	807	1324

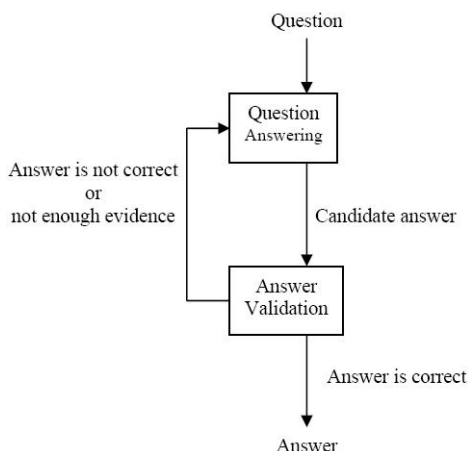


Fig. 2. Decision flow for the Answer Validation

3 Evaluation of the Answer Validation Exercise

The evaluation is based on the detection of the correct answers and only them. There are two reasons for this. First, an answer will be validated if there is enough evidence to affirm its correctness. Figure 2 shows the decision flow that involves an Answer Validation module after searching for candidate answers: In the cases where there is not enough evidence of correctness (according to the AV module), the system must request another candidate answer. Thus, the Answer Validation must focus on detecting that there is enough evidence of the answer correctness.

Second, in a real exploitation environment, there is no balance between correct and incorrect candidate answers, that is to say, a system that validates QA responses does not receive correct and incorrect answers in the same proportion. In fact, the experiences at CLEF during the last years showed that only 23% of all the answers given by all the systems were correct (results for the Spanish as target, see [6]). Although numbers are expected to change, the important thing is that the evaluation of Answer Validation modules must consider the real output of Question Answering systems, which is not balanced. We think that this leads to different development strategies closer to the real problem. Anyway, AVE input must be evaluated with this unbalanced nature.

For these reasons, instead of using an overall accuracy as the evaluation measure, we proposed to use precision (1), recall (2) and a F-measure (3) (harmonic mean) over pairs with entailment value equals to YES. In other words, we proposed to quantify systems ability to detect the pairs with entailment or to detect whether there is enough evidence to accept an answer. If we would had considered the accuracy over all pairs then a baseline AV system that always answers NO (rejects all answers) would obtain an accuracy value of 0.77, which seems too high for evaluation purposes.

$$precision = \frac{|predicted\ as\ YES\ correctly|}{|predicted\ as\ YES|} \quad (1)$$

$$recall = \frac{|predicted\ as\ YES\ correctly|}{|YES\ pairs|} \quad (2)$$

$$F = \frac{2 * recall * precision}{recall + precision} \quad (3)$$

On the other hand, the higher the proportion of YES pairs is, the higher the baselines are. Thus, results must be compared between systems and always taking as reference the baseline of a system that accept all answers (return YES in 100% of cases). Since UNKNOWN pairs are ignored in the evaluation (though they were present in the test collection), the precision formula (1) was modified to ignore the cases were systems assessed a YES value to the UNKNOWN pairs.

4 Results

Eleven groups have participated in seven different languages at this first AVE 2006. Table 2 shows the participant groups and the number of runs they submitted per language. At least two different groups participated in each task (language), so the comparison between different approaches is possible. English and Spanish were the most popular with 11 and 9 runs respectively.

Table 2. Participants and runs per language in AVE 2006

	German	English	Spanish	French	Italian	Dutch	Portuguese	Total
Fernuniversität in Hagen	2							2
Language Computer Corporation		1	1					2
U. Rome “Tor Vergata”		2						2
U. Alicante (Kozareva)	2	2	2	2	2	2	1	13
U. Politecnica de Valencia		1						1
U. Alicante (Ferrández)		2						2
LIMSI-CNRS				1				1
U. Twente	1	2	2	1	1	2	1	10
UNED (Herrera)			2					2
UNED (Rodrigo)			1					1
ITC-irst		1						1
R2D2 project			1					1
Total	5	11	9	4	3	4	2	38

Only 3 of the 12 groups (FUH, LCC and ITC-IRST) have participated in the Question Answering Track showing the chance for new-comers to start developing a single QA

Table 3. AVE 2006 Results for English

System Id	Group	F-measure	Precision	Recall	Techniques
COGEX	LCC	0.4559	0.3261	0.7576	Logic
ZNZ - TV_2	U. Rome	0.4106	0.2838	0.7424	ML
itc-irst	ITC-irst	0.3919	0.3090	0.5354	Lexical, Syntax, Corpus, ML
ZNZ - TV_1	U. Rome	0.3780	0.2707	0.6263	ML
MLEnt_2	U. Alicante	0.3720	0.2487	0.7374	Overlap, Corpus, ML
uaofe_2	U. Alicante	0.3177	0.2040	0.7172	Lexical, Syntax, Logic
MLEnt_1	U. Alicante	0.3174	0.2114	0.6364	Overlap, Logic, ML
uaofe_1	U. Alicante	0.3070	0.2144	0.5404	Lexical, Syntax, Logic
utwente.ta	U. Twente	0.3022	0.3313	0.2778	Syntax, ML
utwente.lcs	U. Twente	0.2759	0.2692	0.2828	Overlap, Paraphrase
100% YES Baseline		0.2742	0.1589	1	
50% YES Baseline		0.2412	0.1589	0.5	
ebisbal	U.P. Valencia	0.075	0.2143	0.0455	ML

Table 4. AVE 2006 Results for French

System Id	Group	F-measure	Precision	Recall	Techniques
MLEnt_2	U. Alicante	0.4693	0.3444	0.7362	Overlap, ML
MLEnt_1	U. Alicante	0.4085	0.3836	0.4369	Overlap, Corpus, ML
100% YES Baseline		0.3741	0.2301	1	
50% YES Baseline		0.3152	0.2301	0.5	
LIRAVE	LIMSI-CNRS	0.1112	0.4327	0.0638	Lexical, Syntax, Paraphrase
utwente.lcs	U. Twente	0.0943	0.4625	0.0525	Overlap

module and, at the same time, opening a place for experienced groups in RTE and Knowledge Representation to apply their research to the QA problem. We expect that in a near future the QA systems will take advantage of this communities working in the kind of reasoning needed for the Answer Validation.

Tables 3-9 show the results for all participant system in each language. Since the number of pairs and the proportion of the YES pairs is different for each language (due to the real submission of the QA systems), results can not be compared between languages. Together with the systems precision, recall and F-measure, two baselines values are shown: the results of a system that always accept all answers (returns YES in 100% of the pairs), and the results of a hypothetical system that returns YES for the 50% of pairs.

In the languages where at least one system reported the use of Logic (Spanish, English and German) the best performing system was one of them. Although the use of Logic does not guarantee a good result, the best systems used it. However, the most extensively used techniques were Machine Learning and overlapping measures between text and hypothesis.

Table 5. AVE 2006 Results for Spanish

System Id	Group	F-measure	Precision	Recall	Techniques
COGEX	LCC	0.6063	0.527	0.7139	Logic
UNED_1	UNED	0.5655	0.467	0.7168	Overlap, ML
UNED_2	UNED	0.5615	0.4652	0.7079	Overlap, ML
NED	UNED	0.5315	0.4364	0.6796	NE recognition
MLEnt_2	U. Alicante	0.5301	0.4065	0.7615	Overlap, ML
R2D2	R2D2 Project	0.4938	0.4387	0.5648	Voting, Overlap, ML
utwente.ta	U. Twente	0.4682	0.4811	0.4560	Syntax, ML
100% YES Baseline		0.4538	0.2935	1	
utwente.lcs	U. Twente	0.4326	0.5507	0.3562	Overlap, Paraphrase
MLEnt_1	U. Alicante	0.4303	0.4748	0.3934	Overlap, Corpus, ML
50% YES Baseline		0.3699	0.2935	0.5	

Table 6. AVE 2006 Results for German

System Id	Group	F-measure	Precision	Recall	Techniques
FUH_1	Fernuniversität in Hagen	0.5420	0.5839	0.5058	Lexical, Syntax, Semantics, Logic, Corpus
FUH_2	Fernuniversität in Hagen	0.5029	0.7293	0.3837	Lexical, Syntax, Semantics, Logic, Corpus, Paraphrase
MLEnt_2	U. Alicante	0.4685	0.3573	0.6802	Overlap, ML
100% YES Baseline		0.3927	0.2443	1	
MLEnt_1	U. Alicante	0.3874	0.4006	0.375	Overlap, Corpus, ML
50% YES Baseline		0.3282	0.2443	0.5	
utwente.lcs	U. Twente	0.1432	0.4	0.0872	Overlap

Table 7. AVE 2006 Results for Dutch

System Id	Group	F-measure	Precision	Recall	Techniques
utwente.ta	U. Twente	0.3871	0.2874	0.5926	Syntax, ML
MLEnt_1	U. Alicante	0.2957	0.189	0.6790	Overlap, Corpus, ML
MLEnt_2	U. Alicante	0.2548	0.1484	0.9012	Overlap, ML
utwente.lcs	U. Twente	0.2201	0.2	0.2469	Overlap, Paraphrase
100% YES Baseline		0.1887	0.1042	1	
50% YES Baseline		0.1725	0.1042	0.5	

Table 8. AVE 2006 Results for Portuguese

System Id	Group	F-measure	Precision	Recall	Techniques
100% YES Baseline		0.3837	0.2374	1	
utwente.lcs	U. Twente	0.3542	0.5783	0.2553	Overlap
50% YES Baseline		0.3219	0.2374	0.5	
MLEnt	U. Alicante	0.1529	0.1904	0.1277	Corpus

Table 9. AVE 2006 Results for Italian

System Id	Group	F-measure	Precision	Recall	Techniques
MLEnt_2	U. Alicante	0.4066	0.2830	0.7219	Overlap, ML
MLEnt_1	U. Alicante	0.3480	0.2164	0.8877	Overlap, Corpus, ML
100% YES Baseline		0.2934	0.1719	1	
50% YES Baseline		0.2558	0.1719	0.5	
utwente.lcs	U. Twente	0.1673	0.3281	0.1123	Overlap

5 Conclusions and Future Work

The starting point for the AVE 2006 was the reformulation of the Answer Validation as a Recognizing Textual Entailment problem, under the assumption that hypotheses can be automatically generated instantiating hypotheses patterns with the QA systems answers. Thus, the collections developed in AVE are specially oriented to the development and evaluation of Answer Validation systems. We have also proposed a methodology for the evaluation in chain with a QA Track. 11 groups have participated with 38 runs in 7 different languages. Systems that reported the use of Logic have obtained the best results in their respective subtasks. Future work aims at reformulating the Answer Validation Exercise for the next campaign and quantifying the gain in performance that the Answer Validation systems introduce in the Question Answering.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Science and Technology within the project R2D2-SyEMBRA. (TIC-2003-07158-C04-02), the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and a PhD grant by UNED. We are grateful to all the people involved in the organization of the QA track (specially to the coordinators at CELCT, Danilo Giampiccolo and Pamela Forner) and to the people that built the patterns for the hypotheses: Juan Feu (Dutch), Petya Osenova (Bulgarian), Christelle Ayache (French), Bodgan Sacaleanu (German) and Diana Santos (Portuguese).

References

1. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The Second PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venezia, Italy (April 2006)
2. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK, pp. 1–8 (April 2005)
3. Herrera, J., Peñas, A., Verdejo, F.: Question Answering Pilot Task at CLEF 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 581–590. Springer, Heidelberg (2005)

4. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M.: The Multiple Language Question Answering Track at CLEF 2003. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 471–486. Springer, Heidelberg (2004)
5. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question Answering Track. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 371–391. Springer, Heidelberg (2005)
6. Peñas, A., Rodrigo, Á., Verdejo, F.: Sparte, a test suite for recognising textual entailment in spanish. In: Gelbukh, A.F. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 275–286. Springer, Heidelberg (2006)
7. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Penas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 Multilingual Question Answering Track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 307–331. Springer, Heidelberg (2006)

Overview of the WiQA Task at CLEF 2006

Valentin Jijkoun and Maarten de Rijke

ISLA, University of Amsterdam
{jijkoun,mdr}@science.uva.nl

Abstract. We describe WiQA 2006, a pilot task aimed at studying question answering using Wikipedia. Going beyond traditional factoid questions, the task considered at WiQA 2006 was to return—given an source page from Wikipedia—to identify snippets from other Wikipedia pages, possibly in languages different from the language of the source page, that add new and important information to the source page, and that do so without repetition.

A total of 7 teams took part, submitting 20 runs. Our main findings are two-fold: (i) while challenging, the tasks considered at WiQA are do-able as participants achieved impressive scores as measured in terms of yield, mean reciprocal rank, and precision, (ii) on the bilingual task, substantially higher scores were achieved than on the monolingual tasks.

1 Introduction

CLEF 2006 featured a pilot on Question Answering Using Wikipedia, or WiQA, for short. The idea to organize a pilot track on QA using Wikipedia builds on several motivations. First, traditionally, people turn to reference works to get answers to their questions. Wikipedia has become one of the largest reference works ever, making it a natural target for question answering systems. Moreover, Wikipedia is a rich mixture of text, link structure, navigational aids, categories, making it extremely appealing for research on text mining and link analysis. And finally, Wikipedia is simply a great resource. It is something we want to work with, and contribute to, both by facilitating access to it, and, as the distinction between readers and authors has become blurred, by creating tools to support the authoring process.

In this overview we first provide a description of the tasks considered and of the evaluation and assessment procedures (Section 2). After that we describe the runs submitted by the participants (Section 3 and detail the results (Section 4). We end with some conclusions (Section 5).

2 Task Definition

The WiQA 2006 task deals with access to Wikipedia’s content, where access is considered both from a point of view and from an author point of view.

As our user model we take the following scenario: a reader or author of a given Wikipedia article (the source page) is interested in collecting information

about the topic of the page that is not yet included in the text, but is relevant and important for the topic, so that it can be used to update the content of the source article. Although the source page is in a specific language (the source language), the reader or author would also be interested in finding information in other languages (the target languages) that he explicitly specifies.

With this user scenario, the task of an automatic system is to locate information snippets in Wikipedia which are:

- outside the given source page,
- in one of the specified target languages,
- substantially new w.r.t. the information contained in the source page, and important for the topic of the source page, in other words, worth including in the content of (the future editions of) the page.

Participants of the WiQA 2006 pilot could take part in two flavors of the task: a monolingual one (where the snippets to be returned are in the language of the source page) and a multilingual (where the snippets to be returned can be in any of the languages of the Wikipedia corpus used at WiQA).

2.1 Document Collections

The data collection used at WiQA 2006 consists of XML-ified dumps of Wikipedia in three language: Dutch, English, and Spanish. The three collections differ greatly in size:

- Dutch: 125,004 articles, 857Mb;
- English: 660,762 articles, 5.9Gb; and
- Spanish: 79,237 articles, 677Mb.

The size of collection and the links structure are important factors for the performance of a system addressing the WiQA 2006 task. Figure 2.1 shows the distribution of the article size and the number of in-links in Wikipedia of the three languages.

The Wikipedia dumps used at WiQA are based on the XML version of the Wikipedia collections [Denoyer and Gallinari(2006)] that include the annotation of the structure of the articles, links between articles, categories, cross-lingual links, etc. For the pilot the annotation of articles was automatically extended with XML markup of sentences and classification of articles into named entity classes (person, location, organization). The classification was done using a set of heuristics that employ the category structure of Wikipedia and the uniform structure of “List” articles (e.g., articles entitled *List of living persons*, *List of physicists*, etc.). The table below shows the distribution of the assigned classes in the collection.

Collection	<i>person</i>	<i>location</i>	<i>organization</i>
English	84,167 (13%)	50,940 (8%)	22,654 (3%)
Spanish	11,009 (14%)	3,980 (5%)	1,292 (2%)
Dutch	10,176 (8%)	7,038 (6%)	1,595 (1%)

We performed a manual assessment of the classes assigned for a random sample of the articles: our heuristic rules resulted in 85% accuracy.

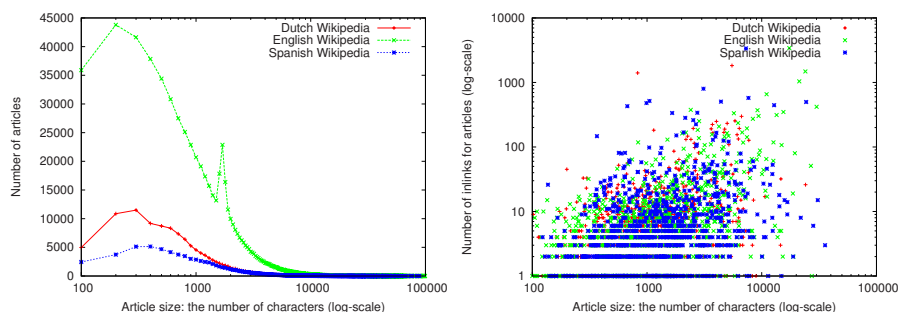


Fig. 1. Distribution of the sizes of the articles and the number of in-links in Dutch, English, and Spanish Wikipedia

2.2 Topics

For each of the three WiQA 2006 languages (Dutch, English, and Spanish) a set of 50 topics correctly tagged as *person*, *location* or *organization* in the XML data collections was released, together with other topics, announced as optional. These optional topics either did not fall into these three categories, or were not tagged correctly in the XML collections. The optional topics could be ignored by systems without penalty. In fact, the submitted runs provided responses for optional topics as well as for the main topics.

When selecting Wikipedia articles as topics, we included articles explicitly marked as stubs, as well as other short and long articles.

In order to create the topics for the English-Dutch bilingual task, 30 topics were selected from the English monolingual topic set and 30 topics from the Dutch monolingual topic set. The bilingual topics were selected so that the corresponding articles are present in Wikipedias for both languages.

The table below shows the number of topics for the four language subtasks.

Task	total	<i>person</i>	<i>location</i>	<i>organization</i>	<i>other</i>
English	65	16	18	16	15
Dutch	60	17	16	17	10
Spanish	67	21	22	18	6
English-Dutch	60	18	16	17	9

In addition to the test topics, a set of 80 (English language) development topics was released.

2.3 Evaluation

Given a source page, automatic systems return a list of short snippets, defined as sequences of at most two sentences from a Wikipedia page. The ranked list of snippets for the topic were manually assessed using the following binary criteria, largely inspired by the TREC 2003 Novelty task [Soboroff and Harman(2003)]:

- *support*: the snippet does indeed come from the specified target Wikipedia article.
- *importance*: the information of the snippet is relevant to the topic of the source Wikipedia article, is in one of the target languages as specified in the topic, and is already present on the page (directly or indirectly) or is interesting and important enough to be included in an updated version of the page.
- *novelty*: the information content of the snippet is not subsumed by the information on the source page
- *non-repetition*: the information content of the snippet is not subsumed by the target snippets higher in the ranking for the given topic

Note that we distinguish between novelty (subsumption by the source page) and non-repetition (subsumption by the higher ranked snippets) in order for the results of the assessment to be re-usable for automatic system evaluation in future: novelty only takes the source page and the snippet into account, while non-repetition is defined on a ranked list of snippets.

One of the purposes of the WiQA pilot task was to experiment with different measures for evaluating the performance of systems. WiQA 2006 used the following main measure for accessing the performance of the systems:

- *yield*: the average (per topic) number of supported, novel, non-repetitive, important target snippets.

We also considered other simple measures:

- *mean reciprocal rank* of the first supported, important, novel, non-repeated snippet, and
- *overall precision*: the percentage of supported, novel, non-repetitive, important snippets among all submitted snippets.

2.4 Assessment

To establish the ground truth, an assessment environment was developed by the track organizers. Assessors were given the following guidelines. For each system and each source article P the ordered list of the returned snippets was be manually assessed with respect to importance, novelty and non-repetition following the procedure below:

1. Each snippet was marked as *supported* or not. To reduce the workload on the assessors, this aspect was checked automatically. Hence, unsupported snippets were excluded from the subsequent assessment.
2. Each snippet was marked as *important* or not, with respect to the topic of the source article. A snippet is important if it contains information that a user of Wikipedia would like to see in P or an author would consider worth to be present in P . Snippets were assessed for importance independently of each other and regardless of whether the important information was already

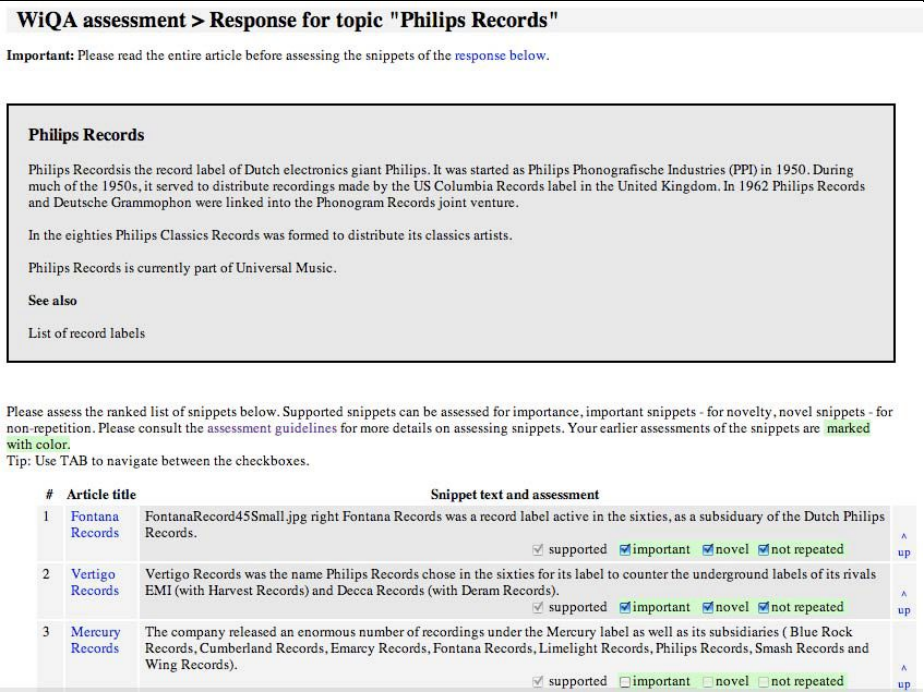


Table 1. Summary of runs submitted to WiQA 2006

Group	Run name	Description
English monolingual task		
LexiClone Inc.	lexiclone	Lexical Cloning method
Universidad Politécnica de València	rfia-bow-en	simple “bag of words” submission
University of Alicante	UA-DLSI-1	Near phrase
	UA-DLSI-2	Near phrase temporal
University of Essex/Limerick	dltg061	Limit of ten snippets per topic
	dltg062	Limit of twenty snippets per topic
University of Amsterdam	uams-linkret-en	Cross-links and IR for snippet ranking
	uams-link-en	Only cross-links for snippet ranking
	uams-ret-en	Only IR for snippet ranking
University of Wolverhampton	WLV-one-old	No coreference, link analysis
	WLV-two	Coreference
	WLV-one	no coreference, version 2
Spanish monolingual task		
Universidad Politécnica de València	rfia-bow-es	simple “bag of words” submission
University of Alicante	UA-DLSI-es	Near phrase
Daedalus consortium	mira-IS-CN-N	InLink sentence retr., rank by novelty
	mira-IP-CN-CN	InLink passage retrieval, combine cosine and novelty in
ranking,		no threshold
Dutch monolingual task		
University of Amsterdam	uams-linkret-nl	Cross-links and IR for snippet ranking
	uams-link-nl	Only cross-links for snippet ranking
	uams-ret-nl	Only IR for snippet ranking
English-Dutch bilingual task		
University of Amsterdam	uams-linkret-ennd	Cross-links and IR for snippet ranking

2.5 Submission

For each task (three monolingual and one bilingual), participating teams were allowed to submit up to three runs. For each topic of a run, the top 20 submitted snippets were manually assessed as described above.

3 Submitted Runs

Table 1 lists the runs submitted to WiQA 2006: 19 for the monolingual task (3 for Dutch, 12 for English and 4 for Spanish) and 1 for the bilingual task (English-Dutch).

Most participating systems used a similar three-step architecture: first, identify snippets relevant to the topic, then estimate their importance, and finally, remove duplicate or near-duplicate snippets. However, there was a lot of variation in the wide range of techniques for addressing individual steps:

- For *identifying relevant snippets* outside the source article, systems used traditional IR (with the title of the source articles as a query), string matching, or made use of the in-links of the article;
- For *estimating the importance of a snippet* the systems employed word overlap, as well as Latent Semantic Analysis, Information Gain or they used the category structure of Wikipedia;
- For *removing redundant snippets* the systems used word overlap, cosine similarity, Information Gain as well as Named Entity identification.

From the text processing perspective, the systems were also very diverse. Participants employed techniques ranging from Named Entity tagging to parsing, logic form identification, coreference resolution and machine translation (using Wikipedia as a training resource for translating proper names between languages). For further details of the individual systems, we refer to the system descriptions in [CLEF 2006 Working Notes(2006)].

In Table 2 we present the aggregate results of the assessment of the runs submitted to WiQA 2006. Columns 3–7 show the following aggregate numbers: total number of snippets (with at most 20 snippets considered per response for a topic); total number of supported snippets; total number of important supported snippets; total number of novel and important supported snippets; and the total number of novel and important supported with repetition.

The results indicate that the task of detecting *important* snippets is a hard one: for most submissions, only 50–60% of the found snippets are judged as important. The performance of the systems for detecting *novel* snippets has a substantially higher range: between 50% and 80% of the found important snippets are judged as novel with respect to the topic article.

4 Results

Table 3 shows the evaluation results for the submitted runs: total yield (for a run, the total number of “perfect” snippets, i.e., supported, important, novel and not repeated), the average yield per topic (only topics with at least one response are considered), the mean reciprocal rank of the first “perfect” snippet and the precision of the systems’ responses.

Clearly, most systems cope well with the pilot task: up to one third of the found snippets are assessed as “perfect” for the English and Spanish monolingual tasks, and up to one half for the Dutch monolingual and the English-Dutch bilingual task. Quite expectedly, the relative ranking of the submitted runs is different for different evaluation measures: as in many complex tasks, the best yield (a recall-oriented measure) does not necessarily lead to the best precision and vice versa.

An interesting aspect of the results is that the performance of the systems differs substantially for the four tasks. This can be due to the fact that the submissions for tasks were assessed by different assessors (native speakers of

Table 2. Results of the assessment of the submitted runs (at most 20 snippets considered per topic)

Run name	Topics with response	Aggregate numbers of snippets				
		total	sup	sup	sup	sup
				imp	imp	imp
				novel	novel	not-rep
English monolingual task: 65 topics						
lexiclone	38	684	676	179	98	79
rfia-bow-en	65	607	607	255	187	173
UA-DLSI-1	64	572	571	277	204	191
UA-DLSI-2	60	489	488	239	173	161
dltg061	65	435	435	226	165	161
dltg062	65	682	682	310	223	194
uams-linkret-en	65	570	570	331	202	191
uams-link-en	65	615	614	353	232	220
uams-ret-en	65	580	580	325	203	193
WLV-one-old	61	473	473	263	219	142
WLV-two	61	526	526	327	280	135
WLV-one	61	473	472	267	221	135
Spanish monolingual task: 67 topics						
rfia-bow-es	62	497	497	198	142	113
UA-DLSI-es	63	501	501	184	149	111
mira-IS-CN-N	67	251	251	127	79	69
mira-IP-CN-CN	67	431	431	155	95	71
Dutch monolingual task: 60 topics						
uams-linkret-nl	60	425	425	301	228	210
uams-link-nl	60	455	455	305	236	228
uams-ret-nl	60	450	450	271	206	192
English-Dutch bilingual task: 60 topics						
uams-linkret-en-nl	60	564	551	456	342	302

the corresponding languages), as well as due to the differences in the sizes and structures of the Wikipedias in these languages. It is worth pointing out that the highest scores were achieved on the English-Dutch bilingual task; this may suggest that different language versions of Wikipedia do indeed present different material on a given topic.

4.1 Inter-annotator Agreement

The definition of the WiQA task is quite complicated and the criteria for snippet assessment may be very subjective. To examine this issue, we arranged the assessments so that a portion of the snippets was assessed by two annotators.

The table below shows the agreement of pairs of assessors on importance judgments: the percent of matching assessments and Cohen’s kappa (κ).

Table 3. Evaluation results for the submitted runs (calculated for top 10 snippets per topic); highest scores per task are given in boldface

Run name	Number of topics with response	Total yield	Average yield	MRR	Precision
English monolingual task: 65 topics					
lexiclone	38	58	1.53	0.31	0.21
rfia-bow-en	65	173	2.66	0.48	0.29
UA-DLSI-1	64	191	2.98	0.53	0.33
UA-DLSI-2	60	158	2.63	0.52	0.32
dltg061	65	160	2.46	0.54	0.37
dltg062	65	152	2.34	0.50	0.33
uams-linkret-en	65	188	2.89	0.52	0.33
uams-link-en	65	220	3.38	0.58	0.36
uams-ret-en	65	191	2.94	0.52	0.33
WLV-one-old	61	142	2.33	0.58	0.30
WLV-two	61	135	2.21	0.59	0.26
WLV-one	61	135	2.21	0.58	0.29
Spanish monolingual task: 67 topics					
rfia-bow-es	62	113	1.82	0.37	0.23
UA-DLSI-es	63	111	1.76	0.36	0.22
mira-IS-CN-N	67	69	1.03	0.30	0.27
mira-IP-CN-CN	67	71	1.06	0.29	0.16
Dutch monolingual task: 60 topics					
uams-linkret-nl	60	210	3.50	0.53	0.49
uams-link-nl	60	228	3.80	0.53	0.50
uams-ret-nl	60	192	3.20	0.45	0.42
English-Dutch bilingual task: 60 topics					
uams-linkret-en-nl	60	302	5.03	0.52	0.54

Assessor pair	Common snippets	Agreement	κ
A,B	91	75%	0.49
C,D	242	86%	0.71
C,B	212	77%	0.52
C,A	77	70%	0.38
D,B	573	72%	0.45
D,E	147	56%	0.13
D,A	46	78%	0.57
F,G	643	73%	0.42

We see that the κ values vary between 0.13 (with an agreement of 56%) to 0.71 (with an agreement of 86%), while most are above 0.4. This indicates a less than perfect correlation between assessors' judgements.

5 Conclusion

We have described the first installment of the WiQA—Question Answering Using Wikipedia—task. Set up as an attempt to take question answering beyond the

traditional factoid format and to one of the most interesting knowledge sources currently available, WiQA had 8 participants who submitted a total of 20 runs for 4 tasks. The results of the pilot are very encouraging. While challenging, the task turned out to be do-able, and in cases several participants managed to achieve impressive yield, MRR, and precision scores. Surprisingly, the highest scores were achieved on the bilingual task.

As to the future of WiQA, as pointed out before we aim to take a close look at our assessments, perhaps add new assessments, and analyse inter-assessor agreement along various dimensions. The WiQA 2006 pilot has shown that it is possible to set up tractable yet challenging information access tasks involving the multilingual Wikipedia corpus—but this was only a first step. In the next edition of the task we would like to put more emphasis on the multilingual aspect of the task and extend the task by allowing systems to locate snippets in a fixed collection of crawled web pages, in addition to Wikipedia articles.

The collections, topics and the assessed runs of the participants of WiQA 2006 are available [WiQA(2006)].

Acknowledgments

We are very grateful to the following people and organizations for helping us with the assessments: José Luis Martínez Fernández and César de Pablo from the Daedalus consortium; Silke Scheible and Bonnie Webber at the University of Edinburgh; Udo Kruschwitz and Richard Sutcliffe at the University of Essex; and Bouke Huurnink and Maarten de Rijke at the University of Amsterdam.

Valentin Jijkoun was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.-065.-120 and 612.000.106. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.-000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

- [CLEF 2006 Working Notes(2006)] CLEF 2006 Working Notes. In: Working Notes for the CLEF 2006 Workshop (2006), http://www.clef-campaign.org/2006/working_working_notes/CLEF2006WN-Contents.html
- [Denoyer and Gallinari(2006)] Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)
- [Soboroff and Harman(2003)] Soboroff, I., Harman, D.: Overview of the TREC 2003 Novelty track. In: Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), NIST, pp. 38–53 (2003)
- [WiQA(2006)] WiQA, Question Answering Using Wikipedia (2006), <http://ilps.science.uva.nl/WiQA/>

Re-ranking Passages with LSA in a Question Answering System

David Tomás and José L. Vicedo

Department of Software and Computing Systems
University of Alicante, Spain
`{dtomas,vicedo}@dlsi.ua.es`

Abstract. As in the previous QA@CLEF competition, two separate groups at the University of Alicante participated this year using different approaches. This paper describes the work of *Alicante 1* group. We have continued with the research line established in the past competition, where the main goal was to obtain a fully data-driven system based on machine learning techniques. The main novelties introduced this year are focused on the information retrieval stage. We developed a module for passage re-ranking based on Latent Semantic Analysis (LSA). This year we extended our participation to monolingual Spanish task and bilingual Spanish-English task.

1 Introduction

This paper focuses on the work of *Alicante 1* group at the University of Alicante (runs *aliv061esen*, *aliv062esen*, *aliv061eses* and *aliv062eses* in [1]). We have continued with the research line established in the past competition [2], where an XML framework was defined with the aim of easily integrating machine learning based modules at every stage of the Question Answering (QA) process.

The main novelties introduced this year are focused on the information retrieval stage. We developed a procedure for passage re-ranking based on Latent Semantic Analysis (LSA) [3]. We employed LSA to compare every question formulated to the system with every passage returned by the search engine in order to re-rank them.

This year we extended our participation from the original monolingual Spanish task to bilingual Spanish-English task. For every task we submitted two runs, one with the baseline system and another with LSA passage re-ranking.

This paper is organized as follows: in Section 2 we describe the system architecture detailing the novelties included this year; Section 3 outlines the runs submitted; Section 4 presents and analyzes the results obtained, paying special attention at the passage re-ranking module; finally, in Section 5 we discuss the conclusions and main challenges for future work.

2 System Description

The main architecture of the system remains unchanged since the last competition [2]. We have a modularized structure where each stage can be isolated in an

independent component so as to easily integrate and substitute new ones in an XML framework. In this way, we want to evolve the system originally described in [4] to a fully data-driven approach, adapting every stage of the system with machine learning techniques.

The system follows the classical three-stage pipeline architecture. Next paragraphs describe each module in detail.

2.1 Question Analysis

This first stage carries out two processes: question classification and query construction. The question classification system is based on Support Vector Machines (SVM), and all the knowledge necessary to classify the question is automatically acquired from the training corpora. This way we can easily train the classifier on different languages by changing the corpus. There is a more detailed description of this module in [5].

On the other hand, the query construction module remains the same for this year competition [4]. We use hand made lexical patters in order to obtain the information needed to build the query for the search engine. We have already developed a fully data-driven keyword detection module [6] to help on query construction although it has not being included in the global QA system yet. In the bilingual Spanish-English task, we just translated the question into English through the SysTran¹ on-line translation service, and applied adapted patterns to obtain the query.

2.2 Information Retrieval

The main novelties introduced this year in our system are focused on this stage. First, we moved from Xapian to Indri search engine due to the computational cost of Xapian's indexing module. Indri offers a powerful and flexible query language, but as our keyword extraction and query construction approach are quite simple, we limited the queries to *ad hoc* retrieval with *query likelihood* ranking approach. The search engine returns the 50 topmost relevant passages with a length of 200 words which are provided to the answer extraction module.

On the other hand, we have developed a re-ranking component for the passages retrieved, trying to weigh them beyond query likelihood criterion. This component is based on LSA. This is a corpus-based statistical method for inducing and representing aspects of the meaning of words and passages reflected in their usage. In LSA a representative sample of text is converted to a term-by-passage matrix in which each cell indicates the frequency with which each term occurs in each passage. After a preliminary information-theoretic weighting of cell entries, the matrix is submitted to Singular Value Decomposition (SVD) and a 100 to 1500 dimensional abstract semantic space is constructed in which each original word and each original passage are represented as vectors. Thus, the similarities derived by LSA are not simple co-occurrence statistics, as the dimension-reduction step constitutes a form of induction that can extract a

¹ <http://www.systransoft.com>

great deal of added information from mutual constraints among a large number of words occurring in a large number of contexts.

In order to test this module, Indri first retrieves the best 1000 passages extracted from the corpora. These passages are employed to construct the term-by-passage matrix with Infomap NLP Software² reducing the original matrix to 100 dimensions by means of SVD. Then, the similarity between the question and every passage is calculated. All the passages are sorted in decreasing order of this value, keeping the 50 best valued passages. This process tries to better rank the passages semantically related to the question, overriding the problem of exact match of keywords from the query.

2.3 Answer Extraction

This module remains almost unchanged since the past competition. The set of possible answers is formed extracting all the n-grams (unigrams, bigrams and trigrams in our experiments) from the relevant passages returned by Indri. First, some heuristics are applied in order to filter non probable answers, like those that contains query terms or do not fit the question type. Next, remaining candidate answers are scored taking into account the sentence where they appear, the frequency of appearance, the distance to the keywords and the size of the answer.

As all these heuristics and features are language independent, the answer extraction module can be either applied to Spanish or English passages without additional changes.

3 Runs Submitted

As in the past three years, we have participated in the monolingual Spanish task. As a novelty, we also took part in the bilingual Spanish-English task. We applied no special cross-lingual techniques as we just automatically translated the questions from Spanish to English to set a monolingual English environment to test the system.

We performed two different runs for every task. The difference between each run is established in the information retrieval stage. The first run represents the baseline experiment, where Indri is employed to retrieve the 50 topmost relevant passages related to the query. The second run introduces the re-ranking module described in Section 2. Indri retrieves the best 1000 passages, which are then re-ranked via LSA to finally select the 50 best ones. The rest of the system remains the same for both runs.

4 Results

This year we submitted four different runs, two for the monolingual Spanish task and two more for the bilingual Spanish-English task.

² <http://infomap-nlp.sourceforge.net>

Table 1. Detailed results for monolingual Spanish and bilingual Spanish-English tasks

Accuracy (%)				
Experiment	Factoid	Definition	Temporally restricted	Overall
aliv061eses	28.08	35.71	0.00	29.47
aliv062eses	26.71	40.48	0.00	29.47
aliv061esen	19.33	22.50	-	20.00
aliv062esen	12.00	27.50	-	15.26

There were a total of 200 questions per run. Table 1 shows the overall and detailed results obtained for factoid, definition and temporally restricted factoid questions. Experiment *aliv061eses* refers to the first run (baseline experiment) in the monolingual Spanish task. Experiment *aliv062eses* refers to the second run (passage re-ranking) in the monolingual Spanish task. Otherwise, experiment *aliv061esen* refers to the first run in the bilingual Spanish-English task, while *aliv062esen* refers to the second run.

Both runs in the monolingual Spanish task present the same overall accuracy (29.47%). Nevertheless, there are differences in the performance depending on the question type. For factoid answers, the first run offers better accuracy, while the second run results offer better precision for definition questions. This tendency is repeated in the bilingual Spanish-English task. So, the re-ranking module seems to perform better for definition questions while decreasing the accuracy of the baseline system over factoid ones.

The experiments in bilingual Spanish-English task reflect a significant loss of performance with respect to the monolingual Spanish runs (20.00% of accuracy in the best run). As for this task we just employed automatic translations of the Spanish questions into English, the noise introduced in this early stage of the question answer process heavily damages the global performance of the system. Thus it is hard to compare the accuracy between monolingual and bilingual tasks. We could override this problem by using perfect hand-made translations only for evaluation purposes.

With respect to list questions, there were ten different questions on every run. We did nothing special to deal with this type of questions in our system. We followed the normal procedure established for every question type, returning up to a maximum of ten answers as indicated in the guidelines. We obtained a P@N of 0.0100, 0, 0.0411 and 0.0200 for *aliv61eses*, *aliv62eses*, *aliv61esen* and *aliv62esen* respectively.

5 Conclusions and Future Work

This year we have continued with the research line established in 2005, where the main goal is to obtain a fully data-driven system based on machine learning techniques. The main novelties this year are focused on the information retrieval stage. We defined a new passage re-ranking module based on LSA that seems to

improve the performance of the system on definition questions, while decreasing the results for factoid ones. This suggests that we could set a good scheme for future work applying the baseline approach for factoid questions while employing re-ranking for definition ones.

Anyway, the results obtained with the re-ranking module were not completely satisfactory. The corpus employed to build the term-by-passage matrix (the 1000 best passages retrieved by Indri on every question) seems inadequate for this task. Probably a bigger amount of text would improve the results. The module is still in a preliminary stage and deserves deeper research.

This year we participated in the bilingual Spanish-English task. The loss of performance with respect to the monolingual Spanish task may be due to the noise introduced by the automatic translation of questions, which makes it a bit difficult to correctly evaluate the system. In spite of that, our results were above the average in the English as target task [1].

As a future work, we plan to better test and tune the individual performance of the re-ranking procedure and continue integrating machine learning based modules in our system, going on with the answer extraction stage. Another challenge for the future is to test the system on new languages.

Acknowledgements

This work has been developed in the framework of the projects CICYT R2D2 (TIC2003-07158-C04) and QALL-ME, which is a 6th Framework Research Programme of the European Union (EU), contract number: IST-FP6-033860.

References

1. Magnini, B., et al.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Working Notes for the CLEF 2006 Workshop, Alicante (Spain) (2006)
2. Tomás, D., Vicedo, J.L., Saiz, M., Izquierdo, R.: An XML-Based System for Spanish Question Answering. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 347–350. Springer, Heidelberg (2006)
3. Landauer, T.K., Foltz, P.W., Laham, D.: Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 259–284 (1998)
4. Vicedo, J.L., Saiz, M., Izquierdo, R.: Does English help Question Answering in Spanish? In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 552–556. Springer, Heidelberg (2005)
5. Bisbal, E., Tomás, D., Vicedo, J.L., Moreno, L.: A Multilingual SVM-Based Question Classification System. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 806–815. Springer, Heidelberg (2005)
6. Tomás, D., Vicedo, J.L., Bisbal, E., Moreno, L.: Automatic Feature Extraction for Question Answering Based on Dissimilarity of Probability Distributions. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) FinTAL 2006. LNCS (LNAI), vol. 4139, pp. 133–140. Springer, Heidelberg (2006)

Question Types Specification for the Use of Specialized Patterns in Prodicos System

E. Desmontils, C. Jacquin, and L. Monceaux

LINA, University of Nantes
2, rue de la Houssiniere, BP92208
F-44322 Nantes CEDEX 3, France

{Emmanuel.Desmontils, Christine.Jacquin, Laura.Monceaux}@univ-nantes.fr

Abstract. We present the second version of the Prodicos query answering system which was developed by the TALN team from the LINA institute. The main improvements made concern in the one hand, the use of external knowledge (Wikipedia) to improve the passage selection step. And on the other hand, the answer extraction step is improved by the determination of four different strategies for locating the answer to a question regarding its type. Afterwards, for the passage selection and answer extraction modules, the evaluation is put forward to justify the results obtained.

1 Introduction

In this paper, we present the second version of the PRODICOS query answering system which was developed by the TALN team from the LINA institute. It was our second participation to the QA@CLEF evaluation campaign. We have participated to the monolingual evaluation task dedicated to the French language. We first present the various modules constituting our system. The question analysis module aims to extract relevant features from questions that will make it possible to guide the passage selection and the answer search. We extract many features from the questions (question category, question type, question focus, answer type and main verb). For this new campaign new features are extracted from the questions in order to improve the passage selection process (named entities, complex terms and dates). We also determine at this step of the process the different strategies used for extracting the answer. They are determined according to the question focus and the question type. We also take into account a new category of question: list category. We then present the selection passage process. The goal of this module is to extract from the journalistic corpora the most relevant passages which answer to the question (ie, the passages which might contain the answer). This year, we present a new strategy applies to definitional queries. We use external knowledge to add information to these kinds of questions. The external source used is the Wikipedia encyclopedia. We expand the question focus (which is either a simple term or a complex term) with pieces of information extracted from wikipedia articles. We then discuss in

detail, the major improvements made on our system concerning the answer extraction module. Afterwards, for the passage extraction module and the answer extraction module the evaluation is put forward to justify the results obtained.

2 Overview of the System Architecture

The Prodicos query answering system is divided into three modules (figure 1):

- question analysis;
- passage extraction (extracts passages which might contain the answer);
- answer extraction (extracts the answer according to the results provided by the previous module).

The modules of the Prodicos's system are based on the use of linguistic knowledge, in particular lexical knowledge coming from the EuroWordnet thesaurus [7] and syntactic knowledge coming from a syntactic chunker which has been developed by our team (by the use of the TreeTagger tool [6]).

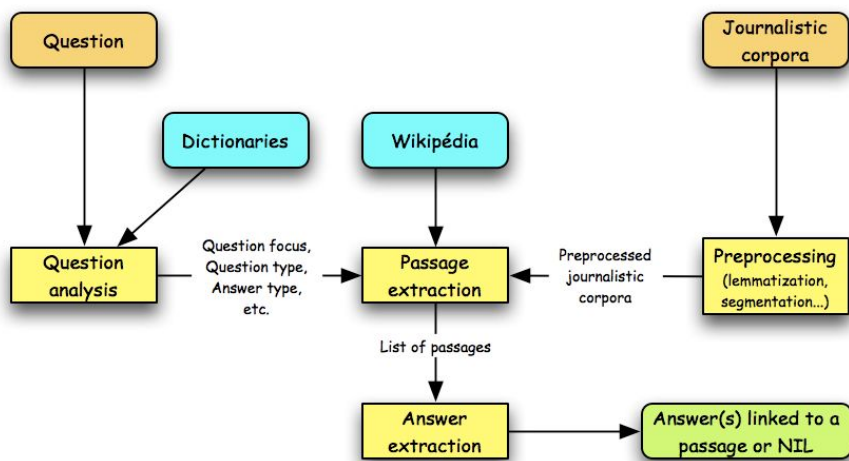


Fig. 1. The global architecture

3 Question Analysis Module

The question analysis module aims to extract relevant features from questions that will make it possible to guide the passage selection and the answer search. We extract many features from the questions [2] [4]: question category, question type, question focus, answer type, main verb, etc.

For this new campaign, new features are extracted from the questions in order to improve the passage selection process (named entities, complex terms

and dates). For example, for the queries: “Qui est Boris Becker ?” (“who is Boris Becker?”) and “Qu’est-ce que l’effet de serre ?” (“what is the greenhouse effect?”), we now consider “Boris Becker” and “effet de serre” as a single entity. We also determine a new feature which is the strategy to use to search the right answer. It is determined according to the question focus and the question type. These strategies are either an entity named strategy, either a numerical entity strategy, either an acronym definition strategy or a pattern-based strategy.

For questions whose answer is of type list, we generate a new feature: the number of answer we are waiting for. In this context, we have three kinds of question, classified according to the number of awaited answers [1].

4 Passage Selection Module

The goal of this module is to extract from the journalistic corpora the most relevant passages which answer to the question (ie, the passages which might contain the answer). A passage is often a sentence excepted for example for a person’s citation which is a set of sentences whose union is regarded as a single passage.

Then, the corpora are indexed by the Lucene search engine¹. The indexing unit used is the passage. For each question, we then build a Lucene request according to the data generated by the question analysis step. The request is built according to a combination of some elements linked with the “or” boolean operator. The elements are: question focus, named entities, main verbs, common nouns, adjectives, dates and other numerical entities [2].

We have a particular strategy for the definitional questions. We use external knowledge to add information to these kinds of questions. The external source used is the Wikipedia encyclopedia. We expand the question focus (which is either a simple term or a complex term) with pieces of information extracted from Wikipedia articles. In this aim, we add to lucene request complex terms or simple terms which belong to the first sentence of the corresponding Wikipedia article (if the question focus exists in the encyclopedia or if the complex term we search the definition is not polysemous). The added complex terms are determined with the help of the TreeTagger tool. For example for the question “Qui est Radovan Karadzic ?” (Who is Radovan Karadzic?), the query sent to lucene search engine is : “Radovan” OR “Karadzic” OR “Radovan Karadzic” OR “homme politique” (“politician”) OR “psychiatre” (“psychiatrist”).

After the CLEF 2006 evaluation campaign, we have made the study, for the definitional queries, of the position of the first passage belonging to the list of returned passage, which contains the right answer (table 1).

The set of Clef evaluation queries comprises 40 definitional queries. For only eight of them, the question focus does not belong to the Wikipedia encyclopedia. This shows that the queries are often general and in this context, the French version of the encyclopedia has a good coverage. Three of them belong to the “wrong tagged” category. For these questions, the TreeTagger tool gives a wrong part-of-speech for the question focus and the consequence is, that the

¹ <http://lucene.apache.org/java/docs/>

Table 1. Passage extraction evaluation for definitionnal queries

	first passage	coef=1 & not first passage	coef < 1	Not present in any passage	Total
Not in Wikipedia	3	3	2	0	8
Wrong tagged	0	1	1	1	3
Not present in corpus	0	0	0	2	2
Other cases	15	5	3	4	27
Total	18	9	6	8	40

passage extraction process does not correctly detect the right passage for the answer. These question focuses are: “Euro Disney”, “Crédit Suisse” and “Javier Clemente”. These named entities are each divided into two simple words and some of these words have a wrong part-of-speech associated to them. Indeed, for French language, some words constituting these named entities are ambiguous and represent either an adjective or a common noun. In table 1, we can see that for 67% of the definitional questions, the passages containing the right answer take the value 1. Moreover, for 83% of them the answer belongs to a passage selected during this step. This are good results and this shows that the use of encyclopedic knowledge helps the selection passage process. The nature of the resource (Wikipedia) is also very interesting because of the recurrent problem for French language to have such kind of resource at one’s disposal. The multilingual property can also be used in a cross-language evaluation context.

5 Answer Extraction

5.1 Global Process

This step comes at the end of our process. After the question analysis and the passage selection, we have to extract correct answers corresponding to questions. To this end, we use on the one hand elements coming from the question analysis like, for instance, the question’s category, the strategy to use it, the number of answers, and so on and, on the other hand, a list of passages selected and evaluated by our previous step according to this question.

The goal of this step is to find the precise answer(s) to a question. An answer is built with the answer itself, the passage used to answer and, a trust value. This ending process can be divided into 4 local steps (figure 2):

1. according to the question’s strategy, the convenient entity extraction module is selected,
2. candidate answers are detected and selected by the previous selected module,
3. answers are evaluated and the answer(s) with the highest trust coefficient is (are) kept,
4. passages where each answer has been found are also associated to the selected answer.

The question’s analysis can give 4 groups of categories which correspond to 4 possible strategies: numerical entities extraction, named entities extraction,

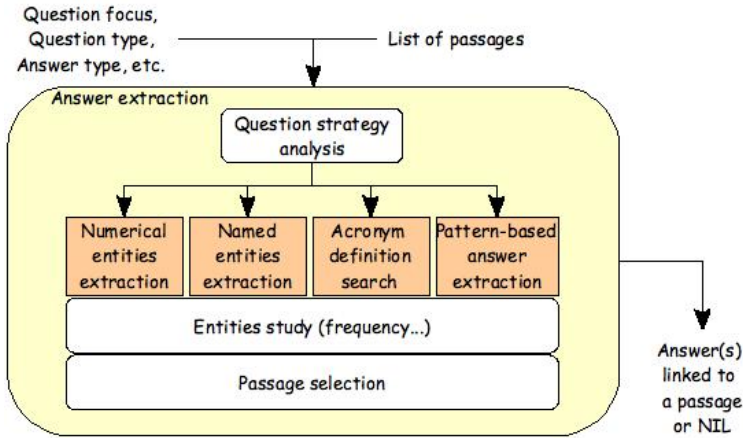


Fig. 2. The answer extraction process

acronym definitions extraction and pattern-based extraction (the default one). Now, we will present processes associated to each strategy and the build of final answers.

5.2 Numerical Entities Extraction

For locating numerical entities, we use a set of dedicated regular expressions. These expressions make it possible to the system to extract numerical information namely: dates, duration, times, periods, ages, financial amounts, lengths, weights, numbers and ratios. It uses the MUC (Message Understanding Conference) categories ("TIMEX" and "NUMEX") to annotate texts.

5.3 Named Entities Extraction

For locating named entities, NEMESIS tool [3] is used. It was developed by our research team. Nemesis is a French proper name recognizer for large-scale information extraction, whose specifications have been elaborated through corpus investigation both in terms of referential categories and graphical structures. The graphical criteria are used to identify proper names and the referential classification to categorize them. The system is a classical one: it is rule-based and uses specialized lexicons without any linguistic preprocessing. Its originality consists on a modular architecture which includes a learning process.

5.4 Acronym Definition Extraction

For acronym's definition search, we use a tool developed by E. Morin [5] based on regular expressions. It detects acronyms and links them to their definition (if it exists). For example, lets take the question: «Qu'est-ce que l' OMS ?» ("What

```

<enonce num="20" num_doc="ATS.941027.0143" coef="1.0">
<texte> Une visite touristique sur ces places ne présente cependant aucun danger , de même qu' un repas à
base de poisson , à condition qu' il soit cuit ou frit , selon une responsable de l' Organisation mondiale de la
santé ( OMS ) . </texte>
<texte-EN> Une visite touristique sur ces places ne présente cependant aucun danger , de même qu' un
repas à base de poisson , à condition qu' il soit cuit ou frit , selon une responsable de l'
<DEF no="1" sigle="OMS">Organisation mondiale de la santé</DEF> ( <SIGLE no="1">OMS</SIGLE> ) .
</texte-EN>
</enonce>

```

Fig. 3. Example of an acronym definition search

means OMS?"). This tool gives results as the one of the figure 3 that shows the analysis the 20th sentence of "ATS.941027.0143"².

5.5 Pattern-Based Answer Extraction

For the pattern-based answer extraction process, we developed our own tool. According to question categories, syntactic patterns were defined in order to extract answer(s) (see figure 4 for an example of a pattern set associated with the question category called "Definition"). These patterns are based on the question focus and makes it possible to the system to extract the answer. Patterns are sorted according to their priority, ie answers extracted by a pattern with an higher priority are considered as better answers than the ones extracted by patterns with a lower priority.

```

<regle>
<Categorie> Definition </Categorie>
<patron>GNRep GNFocus</patron><patron>GPREP GNFocus</patron><patron>GNFocus GNRep</patron>
</regle>

```

Fig. 4. Example of patterns associated to a question category

As a result, for a given question, patterns associated with the question category are applied to all selected passages. Thus, we obtain a set of candidate answers for this question. For example, lets take the question: «Qu'est ce que Hubble ?» ("What is Hubble?"). This question corresponds to the category "Definition" (see figure 4). Pattern-based extraction process gives a set of candidate answers as the one presented in figure 5.

Patterns (syntactic patterns) are based on the noun phrase that contains the focus of the question. Therefore, the first step consists in only selecting passages which could contain the answer and which contain the focus of the question. To apply syntactic patterns, passages are parsed and divided into basic phrases such as noun phrase (GN), adjectival phrase (GA), adverbial phrase (GR), verb phrase (NV), etc. We use a parser which is based on TreeTagger tool for annotating text

² In this example "SIGLE" describes an acronym and "DEF" (with the same identifier "no") its definition.

```

<QUESTION num="0002" nbrep="1" categorie="Definition">
  <REP num="1" doc="LEMONDE95-040794-1.0"> l' objectif </REP>
  <REP num="1" doc="LEMONDE95-023629-1.0"> le télescope </REP>
  <REP num="1" doc="LEMONDE95-033842-1.0"> par le télescope </REP>
  <REP num="1" doc="ATS.940217.0089-1.0">la réparation</REP>
  <REP num="1" doc="LEMONDE95-023628-1.0"> le télescope sa position</REP>...
</QUESTION>

```

Fig. 5. Example of noun phrases extracted using our pattern-based extraction algorithm

with part-of-speech and lemma information. Subsequently, passages are studied to detect the focus noun phrase and to apply each pattern of the question's category.

For the "Definition" category, the pattern strategy gives good results. Nevertheless, for more complex questions, a semantic process could improve the answer search. Indeed, sometimes the build of powerful patterns is a difficult task (such as for the question «Dans quel lieu des massacres de Musulmans ont-ils été commis en 1995 ?» - "Where Moslems were killed in 1995?" -) knowing our patterns are based on the question's focus and the focus is not always easy to find. In addition, the answer type is not always easy to define without semantic information. Another improvement can take into account verb categorization. Indeed, the verb in the question is quite important.

5.6 Answer Selection

When the answer type was been determined by the question analysis step, the process extracts, from the list of passages provided by the previous step, the candidate answers. Named entities, acronym definitions or numerical entities closest to the question focus (if this last is detected) are supported. Indeed, in such cases, the answer is often situated close to the question focus.

The answer selection process depends on the question category. For numerical entities, named entities and acronym definitions, the right answer is the one with the best frequency. This frequency is weighted according to several heuristics such as: the distance (in words) between this answer and the question focus, the presence in the sentence of named entities or dates from the question, etc. For answers extracted by the pattern-based selection, two strategies are used according to the question category:

- the selection of the first selected answer obtained by the first applicable pattern,
- the selection of the most frequent answer (the candidate answer frequency).

Most of the time, the first heuristic is the better one. Indeed, the selected answer is the first one obtained by the first applicable pattern (patterns sorted according to their convenience) and into the first passage (sorted by the passage selection step according to their convenience). Nevertheless, for definitionnal questions such as the question «Qui est Boris Becker ?» ("Who is Boris Becker?") or the

Table 2. Results synthesis

Question type	R	U	X	W	Total
Named entities extraction	23	2	3	53	81
Numerical entities extraction	15	0	4	24	43
Acronym definition search	4	0	0	1	5
Pattern-based answer extraction	16	0	11	44	71
Total	58	2	18	122	200

Table 3. Results comparison between 'R' in the 2005 campaign and 'R' in the 2006 campaign

Question type	2005	2006
Named entities extraction	20.00%	22.86%
Numerical entities extraction	18.42%	26.32%
Acronym definition search	4.35%	65.22%
Pattern-based answer extraction	10.14%	20.29%

first question «Qu'est ce qu'Atlantis ?» ("What is Atlantis?"), we noted that the better strategy is the candidate phrase frequency. Indeed, for this question category where the number of question's terms is low, the passage selection step does not make it possible to the system to select with precision passages containing the answer. Therefore, the frequency-based strategy generally selects the right answer. For example, for the second question, the answer «télescope» is selected for the definitionnal question because of its frequency. Table 5.6 presents all results of our run according to question types.

43 questions (21.5%) was considered as named entities extraction strategy type. Our process finds 15 right answers. 81 questions (40.5%) was considered as named entities extraction strategy type. Our process finds 23 right answers. 71 questions (35.5%) was considered as pattern-based answer search type. Our process finds 16 right answers. 5 questions (2.5%) were analyzed as acronym definition search type. Our process finds 4 right answers. Only "RKA" (for "Agence Spatiale Russe") was not found. This is a specific case. In fact, the definition does not contain the letter "K". Actually, "RKA" is based on the russian definition that does not appear in the corpus. 5 questions awaited a list of answers. In such cases, our process does not produce satisfactory results. For the first "question" (88), a process problem due to the unlimited list causes no answer ! For the second one (92), we give the right answers (with the same associated sentence). For the third one (100), we give only one wrong answer. For the last one (117), we have found 3 of the seven answers, ie (Canada, Russia, Bosnia, Italy, USA, Ukraine, Uruguay) instead of (USA, Canada, Japan, United Kingdom, France, Germany and Italy). The last one was undetected.

In order to evaluate the progress carried out by our system, we tested it with the questions of CLEF05. Regarding the evolution of the results between the data processing sequence 2005 and the data processing sequence 2006, we note a general improvement of the results (we pass from 29 right answers to 55 ones). This progress is not uniform according to the type of the questions. For the named entities type and the numerical entities type, the progress is perceptible

but weak. For these types of questions, we did not modify the basis of the process. We only modified the algorithm for the choice of the right answer. Rather than to take the first answer, we choose to take the most frequent answer in the passages likely to contain the answer. For the pattern-based answer extraction, the progress is more important. Indeed, we have completely implemented again this module by refining the syntactic patterns. To finish, the most spectacular progress is related to the search for definition of acronyms. For this type of question, we now use specialized patterns rather than generic patterns ones to obtain a processing more adapted.

6 Conclusion

In this experiment report, we have studied our second version of the Prodicos system on QA@CLEF2006 question set. The comparison between this version and the first one studied on QA@CLEF2005 is not an easy task. Indeed, question types are quite different. For instance, in the CLEF@2005's session, 21 questions were acronym definition search. Conversely, in the CLEF'2006's session, only 5 questions were acronym definition search. To test the progress of our new system, we have run it on the 2005 data set in order to compare it to our previous version.

For the 2006 session, we notice that the rate of right answers (overall accuracy) are improved of 29% (14.5% at the 2005 session). This result is encouraging (although definitely perfectible). In addition, acronym definition questions are less numerous while our tool is more powerful. The rate of good answers is definitely low but some improvements were made. For instance, the pattern-based answer extraction found 16 answers whereas it found only 2 answers last year. In addition, in this process, 11 answers are inexact (in the set "X"). As a result, it is obvious that the syntactic parser has to be improved to reduce the set "X" and, consequently, to increase the set "R".

Positive points are: (1) a good study of the question type, (2) a correct passage search and (3) an improvement of the answer extraction process. Nevertheless, some improvements have to be done concerning: (1) the question focus identification, (2) the use of semantic resources in French language for all process steps and (3) the answers extraction processes. Furthermore, we have to improve our French semantic ressource. Indeed, EuroWordnet in its French version has some defaults like the lack of definitions for concepts, relations between some concepts are unavailable, etc.

The progress between the data processing sequence 2005 and the data processing sequence 2006 is thus perceptible. We thus will carry our effort on the following points:

- we will try to improve the choice of the answer. Indeed, the two strategies tested to recover the answer are not completely satisfactory. It seems that an intermediate strategy, according to the type of the question, could give better results;
- the syntactic patterns used have to be refined and extended for the search of the named entities;

- our current analyzer has limited performances, it would not be useless to improve the parser (by using patterns for example);
- finally, we will try to better take into account the concept of time on all the levels of the process (question analysis, passage extraction...).

References

1. Desmontils, E., Jacquin, C., Monceaux, L.: Prodicos experiment feedback for QA@CLEF2006. In: Working Notes for the CLEF 2006 Workshop, 20-22, September Alicante, Spain (2006)
2. Monceaux, L., Jacquin, C., Desmontils, E.: The query answering system Prodicos. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 527–534. Springer, Heidelberg (2006)
3. Fourour, N.: Identification et catégorisation automatiques des entités nommées dans les textes français, PhD Thesis, Université de Nantes, LINA (2004)
4. Monceaux, L.: Adaptation du niveau d'analyse des interventions dans un dialogue : application à un système de question-réponse, PhD Thesis, Paris Sud, Orsay (2003)
5. Morin, E.: Extraction de liens sémantiques entre termes à partir de corpus de textes techniques, PhD Thesis, Université de Nantes, LINA (Décember 1999), <http://www.sciences.univ-nantes.fr/info/perso/permanents/morin/article/morin-these99.pdf>
6. Schmid, H.: Improvements in Part-of-Speech Tagging with an Application To German. In: Armstrong, S., et al. (eds.) Natural Language Processing Using Very Large Corpora, Kluwer Academic Publisher, Dordrecht (1999)
7. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic. editor Networks Piek Vossen, university of Amsterdam (1998)

Answer Translation: An Alternative Approach to Cross-Lingual Question Answering

Johan Bos¹ and Malvina Nissim²

¹ Linguistic Computing Laboratory
Department of Computer Science
University of Rome “La Sapienza”, Italy
`bos@di.uniroma1.it`

² Dept. of Language and Oriental Studies
Faculty of Humanities
University of Bologna, Italy
`malvina.nissim@unibo.it`

Abstract. We approach cross-lingual question answering by using a mono-lingual QA system for the source language and by translating resulting answers into the target language. As far as we are aware, this is the first cross-lingual QA system in the history of CLEF that uses this method—almost without exception, cross-lingual QA systems use translation of the question or query terms instead. We demonstrate the feasibility of our alternative approach by using a mono-lingual QA system for English, and translating answers and finding appropriate documents in Italian and Dutch. For factoid and definition questions, we achieve overall accuracy scores ranging from 13% (EN→NL) to 17% (EN→IT) and lenient accuracy figures from 19% (EN→NL) to 25% (EN→IT). The advantage of this strategy to cross-lingual QA is that translation of answers is easier than translating questions—the disadvantage is that answers might be missing from the source corpus and additional effort is required for finding supporting documents of the target language.

1 Introduction

Question Answering (QA) is the task of providing an exact answer (instead of a document) to a question formulated in natural language. Cross-lingual QA is concerned with providing an answer in one language (the target language) to a question posed in a different language (the source language). Most systems tackle the cross-lingual problem by translating the question (or query terms) posed in the source language in the target language, and then using a mono-lingual QA system developed for the target language for retrieving an answer.

Surprisingly little attention has been given to an alternative approach: translating the answer, instead of the question. The main advantage of this method is that answers are easier to translate than questions, due to their simpler syntactic structure. In fact, some types of answers (such as date expressions or names of persons) hardly need a translation at all. In addition, finding a document

in the target language supporting the answer (as prescribed in the QA@CLEF exercise) is feasible: using either the translated question or keywords thereof and the translated answer, standard document retrieving tools can be used to find a document. This approach can work provided that the source and target language documents cover the same material, otherwise all bets are off.

In the context of CLEF, we tested this approach relying on the fact that the documents in the collection of newswire articles come from the same time period, and hence are likely to cover approximately the same topics. We ran our experiments for two language pairs, with English as source language, and Dutch (EN→NL) and Italian (EN→IT), as target language, respectively. We used an existing mono-lingual QA system for English and an off-the-shelf machine translation tool for translating the answers. In this report we describe and evaluate our method in the context of CLEF, and discuss the feasibility of this approach in general.

2 Related Work

Almost without exception, the method used in cross-lingual QA for crossing the language barrier is translating the question of the source language into the target language, and then using a mono-lingual QA system for the target language. A variant on this method is translating the query terms derived from the analysis of the source question into terms of the target language. Both methods apply the translation step very early in the pipeline of QA components.

To get an accurate overview of all the approaches, we carried out a survey on the methods used in all the editions of cross-lingual QA@CLEF, based on information gathered from the working notes of CLEF 2003 [7], CLEF 2004 [10], CLEF 2005 [8] and CLEF 2006 [9]. In these four years of CLEF we counted 44 systems participating in 68 (26 different ones) cross-lingual tasks. Specifically, the tasks considered were BU→EN (3x), BU→FR, DE→EN (5x), DE→FR, EN→DE (4x), EN→ES (6x), EN→IT (2x), EN→DN, EN→FR (5x), EN→NL (3x), EN→PT (2x), ES→EN (4x), ES→FR, ES→PT, FI→EN (2x), FR→EN (11x), FR→ES, IT→EN (3x), IT→ES, IT→FR (2x), NL→EN, NL→FR, PT→EN, PT→FR (3x), IN→EN (2x), and RO→EN.

In the majority of cases (63 of 68), the cross-lingual problem was addressed by using off-the-shelf translation software (such as Babelfish, Systran, Reverso, FreeTrans, WorldLingo, Transtool) to translate the question or the keywords of the query of the source language into the target language, followed by processing the translated question or query using a mono-lingual QA system.

The alternative method we propose applies the translation step at the last stage of the QA pipeline. Instead of translating the question, a mono-lingual QA system for the source language produces an answer which is subsequently translated in the target language. In the CLEF-2006 campaign we applied our method to the language pairs EN→IT and EN→NL. As far as we are aware, this is the first time, in the context of QA@CLEF, that such an approach was

implemented and evaluated. A similar approach to ours was also presented at CLEF 2006 by LCC [3]. They ran a QA system for the source language on an automatically translated (from target to source language) document collection, and then aligned the answer found in the source language document with that of the answer and document in the target language. This was done for three language pairs: EN→FR, EN→ES, and EN→PT. Although our method is similar, it does not require the effort to translate the entire document collection, and focuses on just translating the obtained answer.

3 Method

Our cross-lingual QA system is based on a mono-lingual QA system for English extended with an answer translation module. It deals with factoid (including list questions) and definition questions, but it uses two different streams of processing for these two types of questions. The first component in the pipeline, Question Analysis, deals with all types of questions. Then, when the question turns out to be of type factoid, Document Analysis, Answer Extraction, and Answer Translation (and document support) will follow. Answers to definition questions are directly searched in the target language corpora (see Figure 1 and Section 3.5). What follows is a more detailed description of each component. An example of the different data structures of the system is shown in Fig. 2.

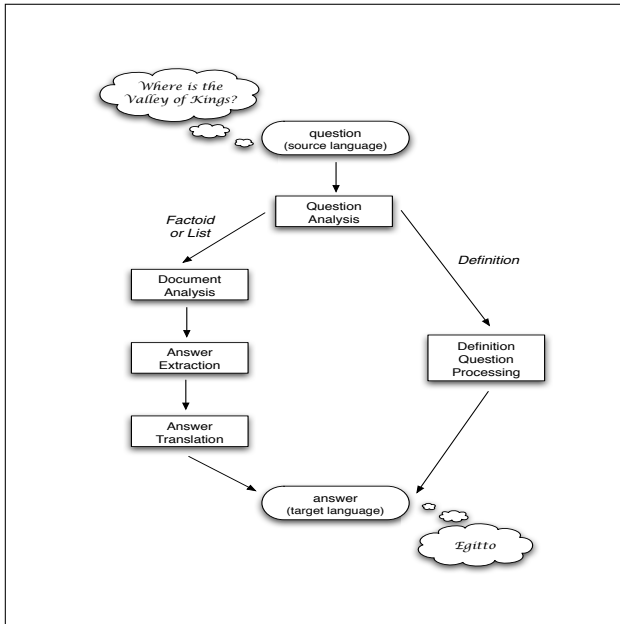


Fig. 1. Simplified architecture of the QA system

3.1 Question Analysis

The (English) question is tokenised and parsed with a wide-coverage parser based on Combinatory Categorical Grammar (CCG). We use the parser of Clark & Curran [4]. On the basis of the output of the parser, a CCG-derivation, we build a semantic representation in the form of a Discourse Representation Structure (DRS), closely following Discourse Representation Theory [5]. This is done using the semantic construction method described in [1,2]. The Question-DRS is the basis for generating four other sources of information required later in the question answering process: an expected answer type; a query for document retrieval; the answer cardinality; and background knowledge for finding appropriate answers.

We distinguish 14 main expected answer types which are further divided into subtypes. The main types are definition, description, attribute, numeric, measure, time, location, address, name, language, creation, instance, kind, and part. The answer cardinality denotes a range expressed by an ordered pair of two numbers, the first indicating the minimal number of answers expected, the second the maximal number of answers (or 0 if unspecified). For instance, 3–3 indicates that exactly three answers are expected, 2–0 means at least two answers. Background knowledge is a list of axioms related to the question—it is information gathered from WordNet or other lexical resources.

3.2 Document Analysis

In order to maximise the chance of finding an answer in the source language (English), we extended the English CLEF document collection with documents from the Acquaint corpus. All documents were pre-processed by applying sentence splitting and tokenisation and dividing them into smaller documents of two sentences each (taking a sliding window, so each sentence will appear in two mini-documents). These mini-documents were indexed with the Indri information retrieval tools (we used version 2.2, see <http://www.lemurproject.org/indri/>). The query generated by Question Analysis (see Section 3.1) is used to retrieve the best 1,000 mini-documents, again with the use of Indri.

Using the same wide-coverage parser as for parsing the question, all retrieved documents are parsed and for each of them a Discourse Representation Structure (DRS) is generated. The parser also performs basic named entity recognition for dates, locations, persons, and organisations. This information is used to assign the right semantic type to discourse referents in the DRS.

3.3 Answer Extraction

Given the DRS of the question (the Q-DRS), and a set of DRSs of the retrieved documents (the A-DRSs), we match each A-DRS with the Q-DRS to find a potential answer. This process proceeds as follows: if the A-DRS contains a discourse referent of the expected answer type (see 3.1) matching will commence attempting to identify the semantic structure in the Q-DRS with that of the A-DRS. The result is a score between 0 and 1 indicating the amount of semantic

material that could be matched. The background knowledge (such as hyponyms from WordNet) generated by the Question Analysis (see 3.1) is used to assist in the matching. All retrieved answers are re-ranked on the basis of the match-score and frequency.

3.4 Answer Translation

The answer obtained for the source language is now translated into the target language by means of the Babelfish off-the-shelf machine translation tool. However, this is not done for all types of answers: we refrain from translating answers that are person names or titles of creative works. Indeed, person names are often not translated across languages. Names of creative works instead can be, but the machine translation software that we use does not perform well enough on some examples in our training data, so that we decided to leave these untranslated, too. In addition, titles of creative works often don't get a literal translation (a case in point is Woody Allen's "Annie Hall", which is translated in Italian as "Io e Annie") so a more sophisticated translation strategy would be required.

Document Support. Given the answer in the target language, we need to find a supporting document from the target language collection (as prescribed by the QA@CLEF exercise). Hence, we also translate the original question, independently of the answer found for it. This is used to construct another query for retrieving a document from the target language collection that contains the translated answer and as many as possible terms from the translated question. (As with the source language documents, these are indexed and retrieved using Indri.)

3.5 Definition Question Processing

We did not put much effort in dealing with definition questions. We simply adopted a basic pattern matching technique directly on the target language documents. Once a question is identified as a definition question (see Figure 1), all non-content words are removed from the question (wh-words, the copula, articles, punctuation, etc.). What is left is usually just the topic of the definition question. For instance, for the English question "Who is Radovan Karadzic?" we derive the topic "Radovan Karadzic".

Given a topic, we obtain interesting information about it expressed in the target language. We do this by searching an off-line dump of the Dutch and Italian Wikipedia pages for sentences mentioning the target. From this we generate an Indri query selecting all (one-sentence) documents containing the topic, and a combination of the terms found in Wikipedia (removing stop words). This yields a list of documents in the target language.

The last step is based on template matching using regular expressions to extract the relevant clauses of a sentence that could count as an answer to the question. There are only a few templates, but we use different ones for the different sub-types of definition questions (the Question Analysis component distinguishes between three sub-types of definition questions: person, organisation,

and thing). We developed a few general patterns tested on definition questions from previous QA challenges, which were both valid for Dutch and Italian:

```

person:      /(^| )(\,|de|i|l|la|i|gli|l|') (.) $topic /i
person:      /$topic \( (\d+) \)/i
organisation: /$topic \((\[^\])+\)\)/i
thing:       /$topic , (\[,]+) , /
thing:       /$topic \((\[^\])+\)\)/i

```

As an example of what these simple patterns can achieve, consider the following sentences (extracted answers are type-set in bold-face):

LASTAMPA94-041510 Il **leader** Radovan Karadzic non era reperibile.
 LASTAMPA94-042168 Intanto il **leader serbo-bosniaco** Radovan Karadzic e
 il comandante in capo delle forze serbe, gen.

4 Evaluation

4.1 Performance at CLEF-2006

The question analysis component performed fairly well. Only 2 of 200 questions of the EN→IT set, and 7 of EN→NL set could not be parsed. For 189 questions of the EN→IT set, and 177 of the EN→NL set the system determined an expected answer type. This already shows that our system had more difficulties with the EN→NL set, which is reflected in the overall scores.

Relatively many questions did not have an answer in the English collection, or at least our system failed to find one. For 41 of the 200 EN→NL questions and for 43 of the 200 EN→IT questions no English answer (correct or incorrect) was found. Answer translation introduced only few errors (see Section 4.2), but finding a supporting document proved harder than we had hoped. So several correct answers were associated with wrong documents and therefore judged as “unsupported” (see below).

We submitted four runs—two for the EN→NL task, and two for the EN→IT task. The first runs contained one answer for factoid questions and up to ten for definition and list questions. The second runs contained up to ten answers for each type of question. We used this strategy because it was unclear, at the time of submission, what kind of evaluation would be used. So the number one runs would perform better on an accuracy score, and the number two runs better on scores based on the average mean reciprocal rank (MRR).

Eventually, both accuracy and MRR were used in the evaluation (see [6]). The results for definition and factoid questions are shown in Table 1, and the figures for the list questions in Table 2. As the tables illustrate, the scores for accuracy are the same for each run, which means that providing more than one answer for a factoid question was not punished. As expected, we achieved better scores for MRR on the number two runs. It is surprising that we did so well on definition questions, given that we paid very little effort in dealing with them.

Input Question: [0040] What year did Titanic sink?	
Q-DRS:	x1 year(x1) ? x2 x3 named(x3,titanic,nam) sink(x2) agent(x2,x3) rel(x2,x1)
Answer Type: [tim:yea]	
Cardinality: 1-1	
English Query: #filreq(Titanic #weight(1 Titanic 3 sink))	
English Answer: 1912	
English Context: [APW19981105.0654] Speaking of bad guys: Cartoon News reprints a bunch of political cartoons published after the <u>Titanic sank in 1912</u> .	
A-DRS:	x0 x1 x2 x3 speak(x1) agent(x1,x0) bad(x2) guy(x2) proposition(x3) :(x2,x3) x4 x5 x6 x6 x7 x8 x9 named(x4,cartoon_news,org) bunch(x5) political(x6) publish(x7) patient(x7,x6) proposition(x8) after(x7,x8) x10 x11 x12 named(x10,titanic,nam) timex(x12)=1912-XX-XX sink(x11) agent(x11,x10) in(x11,x12) cartoon(x6) of(x5,x6) reprint(x9) agent(x9,x4) patient(x9,x5) of(x1,x2)
Italian Answer: 1912	
Italian Query: #filreq(1912 #combine(1912 anno Titanic affondato))	
Italian Context: [AGZ.941120.0028] I cantieri di Belfast hanno confermato di aver ricevuto la commessa più prestigiosa dopo la costruzione del "Titanic": la messa in mare del "Titanic 2". Una copia perfetta del fantasmagorico transatlantico affondato nell' oceano nel 1912 durante il viaggio inaugurale è stata ordinata da un consorzio giapponese presso la ditta "Harland and Wolff" e sarà completata entro il 1999.	

Fig. 2. System input and output for a factoid question in the EN→IT task

Table 1. Number of right (R), wrong (W), inexact (X), unsupported (U) answers, accuracy measure of first answers (factoid, definition, overall, and lenient), and mean reciprocal rank score (MRR) for all four submitted CLEF 2006 runs

Run	R	W	X	U	Accuracy				MRR
					Fact.	Acc.	Def.	Acc.	Overall Lenient
EN→IT 1	32	141	4	11	15.3%	24.4%	17.0%	25.0%	0.18
EN→IT 2	32	141	4	11	15.3%	24.4%	17.0%	25.0%	0.20
EN→NL 1	25	150	6	3	11.6%	20.5%	13.4%	18.5%	0.14
EN→NL 2	25	149	7	3	11.6%	20.5%	13.4%	19.0%	0.15

Table 2. Number of right (R), wrong (W), inexact (X), unsupported (U), unassessed (I) answers, P@N (percentage of correct answers in returned set) score for list questions, for all submitted CLEF 2006 runs

Run	R	W	X	U	I	P@N
EN→IT 1	12	63	1	6	0	0.10
EN→IT 2	15	70	1	7	0	0.15
EN→NL 1	6	29	0	7	24	0.18
EN→NL 2	9	40	0	5	42	0.15

4.2 Answer Translation Accuracy

Recall that depending on the kind of expected answer type, answers were translated into the target language or not. We did not translate names of persons, nor titles of creative works. Whereas this strategy worked out well for person names, it didn't for names of creative works. For instance, for question "0061 What book did Salman Rushdie write?", we found a correct answer "Satanic Verses" in the English documents, but this was not found in the Italian document collection, because the Italian translation is "Versetti Satanic". Babelfish wouldn't have helped us here either, as it translates the title into "Verses Satanic".

Table 3. Answer translation accuracy, divided over answer types (EN→NL). Quantified over the first translated answer, disregarding whether the answer was correct or not, but only for questions to which a correct expected answer type was assigned.

Answer Type	Correct	Wrong	Accuracy
location	30	0	100%
numeric	9	2	82%
measure	4	2	67%
instance (names)	14	5	74%
time	15	0	100%
other	7	1	88%

Overall, answer translation introduced little errors as many answers are easy to translate (Table 3). Locations are usually correctly translated by Babelfish, and so are time expressions. Difficulties sometimes arise for numeric expressions. For instance, for the question “0084 How many times did Jackie Stewart become world champion?” we found the correct English answer “three-time”. However, this was wrongly translated in “drie-tijd”, and obviously not found in the Dutch corpus. Similarly, many answers for measurements are expressed in the English corpus using the imperial system, whereas the metric system is used in the Dutch and Italian corpora. Finally, some things are just hard to translate. Although we found a reasonably correct answer for “0136 What music does Offspring play?”, namely “rock”, this was translated in Dutch as “rots”. In itself a correct translation, but for the wrong context.

5 Discussion

What can we say about the answer-translation strategy to cross-lingual question answering, and how can we improve it? Generally speaking, we believe it is a promising approach, but success depends on improvement in three areas: translation itself, source document coverage, and target document support.

Translation. Generally, answers are easier to translate than questions since they are syntactically less complex. For many answer types, word sense ambiguity can cause erroneous translations. In addition, there are issues specific to certain answer types. Table 4 summarises and exemplifies some of them.

Table 4. Aspects of answer translation in different answer types

Answer Type Aspect		Source	Target
numeric	punctuation	100,000 [EN]	100.000 [IT]
measure	unit conversion	26 miles [EN]	42 km [NL]
time	formatting	the 3rd of January, 1982 [EN]	3 Januari 1982 [NL]
location	spelling	London [EN]	Londra [IT]
		New York [IT]	New York [IT]
creation	translation	The Office [EN]	The Office [IT]
		Annie Hall [EN]	Io e Annie [IT]

As Table 4 shows, each answer type comes with specific problems in translation. For instance, for creative works and locations, sometimes terms are translated and sometimes they aren’t. It seems that many difficulties in translating answers might be addressed by dedicated translation strategies and look-up tables for answers expressing measure terms and titles of creative works. In general, knowing the answer types is a great advantage for obtaining a high-quality translation.

Source Document Coverage. One problem that came to the surface was the omission of the answer in the source language document collection. There is only one way to deal with this situation, and that is getting a larger pool of documents. One option is to use the web for finding the answer in the source language.

Target Document Support. Another problem that arised was finding a supporting document for the target language. The system can certainly be improved with respect to this point: it currently only takes the translated question as additional information to find a target language document. This is interesting, as the machine translated question need not be grammatically perfect to find a correct document. One way to improve this is to translate the context found for the source language as well, and use it in addition to retrieve a target language document.

References

1. Bos, J., Clark, S., Steedman, M., Curran, J.R., Hockenmaier, J.: Wide-Coverage Semantic Representations from a CCG Parser. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING '04), Geneva, Switzerland (2004)
2. Bos, J.: Towards wide-coverage semantic interpretation. In: Proceedings of Sixth International Workshop on Computational Semantics IWCS-6, pp. 42–53 (2005)
3. Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J., Moldovan, D.: LCC's PowerAnswer at QA@CLEF 2006. In: Peters (ed.) Working Notes for the CLEF 2006, September 20–22, Workshop, Alicante, Spain (2006)
4. Clark, S., Curran, J.R.: Parsing the WSJ using CCG and Log-Linear Models. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04), Barcelona, Spain (2004)
5. Kamp, H., Reyle, U.: From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. Kluwer, Dordrecht (1993)
6. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Osenova, P., Penas, A., Jijkoun, V., Sacaleanu, B., Rocha, P., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Peters (ed.) Working Notes for the CLEF 2006 Workshop, September 20–22, Alicante, Spain (2006)
7. Peters, C.: Results of the CLEF 2003 Cross-Language System Evaluation Campaign. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 21–22. Springer, Heidelberg (2004)
8. Peters, C.: Results of the CLEF 2005 Cross-Language System Evaluation Campaign. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 21–23. Springer, Heidelberg (2006)
9. Peters, C.: Results of the CLEF 2006 Cross-Language System Evaluation Campaign. In: Working Notes for the CLEF 2006 Workshop, September 20–22, Alicante, Spain (2006)
10. Peters, C., Borri, F.: Results of the CLEF 2004 Cross-Language System Evaluation Campaign. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 15–17. Springer, Heidelberg (2005)

Priberam's Question Answering System in a Cross-Language Environment

Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes,
Cláudia Pinto, and Daniel Vidal

Priberam Informática
Alameda D. Afonso Henriques, 41 - 2º Esq.
1000-123 Lisboa, Portugal
{ach, hgf, atm, amm, prm, cp, dpv}@priberam.pt

Abstract. Following last year's participation in the monolingual question answering (QA) track of CLEF, where Priberam's QA system has achieved state-of-the-art results, this year we decided to take part in both Portuguese and Spanish monolingual tasks, as well as in two bilingual (Portuguese-Spanish and Spanish-Portuguese) tasks of QA@CLEF. This paper describes the improvements and changes implemented in Priberam's QA system since last CLEF participation, detailing the work involved in its cross-lingual extension and discussing the results of the runs submitted to evaluation.

1 Introduction

Priberam took part in the 2005 CLEF campaign, introducing its QA system for Portuguese in the monolingual task. This year, encouraged by last year's results [1], we decided to participate in the same track, but we extended the participation to the Spanish monolingual and bilingual (Portuguese-Spanish and Spanish-Portuguese) tasks.

The architecture of the QA system remains roughly the same [2,3]. It relies on previous work done for a multilingual semantic search engine developed in the framework of TRUST¹, where Priberam was responsible for the Portuguese module. Unlike TRUST, however, which used a third party indexation engine, the system is based on the indexing technology of LegiX, Priberam's legal information tool². The procedure is as follows: after the question is submitted, it is categorised according to a question typology. An internal query retrieves a set of potentially relevant documents, containing a list of sentences related with the question. Sentences are weighted according to their semantic relevance and similarity with the question. Next, through specific answer patterns, these

¹ TRUST – Text Retrieval Using Semantic Technologies – was a European Commission co-financed project (IST-1999-56416), whose aim was the development of a multilingual semantic search engine capable of processing and answering natural language questions in English, French, Italian, Polish and Portuguese.

² See <http://www.legix.pt> for more information.

sentences are examined once again and the parts containing possible answers are extracted and weighted. Finally, a single answer is chosen among all candidates.

This year, we focused on the handling of temporally restricted questions, the addition of another language, and the adaptation of the system to a cross-language environment. Priberam relied on the scalability of the system's architecture to cope with the inclusion of another language module, Spanish. Currently, two M-CAST³ partners, the University of Economics, Prague (UEP) and TiP, are also using this framework to develop Czech and Polish language modules. The purpose of this year's participation in QA@CLEF was to evaluate the language independence of the system, and its performance in a bilingual context.

The next section gives an overview of the work done for the inclusion of another language, describing the development of the Spanish module. Section 3 addresses the various improvements in Priberam's QA system architecture and depicts the cross-language scenario. Section 4 discusses both the monolingual and bilingual results of the system at this year's QA@CLEF campaign, and section 5 concludes with future guidelines.

2 Addition of Spanish

Taking advantage of the company's natural language processing (NLP) technology and workbench [3,4], it was somewhat straightforward to build a new language module for Spanish, similar to the Portuguese one. The QA system architecture was designed to be language independent, and by using the same software tools, like Priberam's SintaGest, new language modules can easily be implemented and tested. This means that only the language resources (lexicon, thesaurus, ontology, QA patterns) have to be adapted or imported to be in conformity with the existing NLP tools.

The Spanish ontology was added to the common multilingual ontology, in a joint work of Priberam and Synapse Développement. As detailed in [2], this multilingual taxonomy groups over 160 000 words and expressions through their conceptual domains organised in a four level tree with 3 387 terminal nodes.

Priberam started by acquiring a Spanish lexicon and converting it to the same format and specifications of the Portuguese one. After loading the existent lexical information in a database, we had to establish equivalencies between POS categories, uniform them and classify all the lexical entries, so the Spanish lexicon could be used with the tools that were used to build the Portuguese module. New entries in the lexicon were also inserted, mainly proper nouns, such as toponyms and anthroponyms. This was particularly important for the recognition of named entities (NEs).

³ M-CAST – Multilingual Content Aggregation System based on TRUST Search Engine – is an European Commission co-financed project (EDC 22249 M-CAST), whose aim is the development of a multilingual infrastructure enabling content producers to access, search and integrate the assets of large multilingual text (and multimedia) collections, such as internet libraries, resources of publishing houses, press agencies and scientific databases (<http://www.m-cast.infovide.pl>).

Unlike the Portuguese lexicon, the Spanish one does not contain any semantic features, nor sense definitions connected to the ontology levels. The semantic classification has already been started but we still have to define senses for each entry in the lexicon and link each one of them to the levels of the ontology. This future work in the Spanish lexicon will hopefully lead to similar results of both language modules. There is also still no thesaurus for Spanish, which may influence the retrieval of documents and sentences that contain synonyms of the question's keywords.

Some relations between words were semi-automatically established, such as derivation relations (e.g. *caracterizar/caracterización*), names of toponyms and their gentiles (e.g. *Australia/australiano*), currencies (e.g. *Grecia/euro*) and cities (e.g. *España/Madrid*). Other relations, such as hyperonymy, hyponymy, antonymy and synonymy, still need to be implemented in the Spanish lexicon and improved in the Portuguese one.

The system's overall design remained unchanged for Spanish; the question patterns, answer patterns and question answering patterns [3] were adapted to fit the new language module. The typology of the 86 question categories was not altered, since it is language independent. Due to the syntactical similarities between the two Romance languages, many of the Portuguese patterns remained applicable.

Like in Portuguese, Spanish morphological disambiguation is done in two stages: first, the contextual rules defined in SintaGest are applied; then, remaining ambiguities are suppressed with a statistical POS tagger based on a second-order hidden Markov model [5,6]. For training, we used a corpus previously disambiguated with SVMTool⁴ [7].

3 Cross-Language and Architecture Improvements

The five main steps of Priberam's QA system are the *indexing process*, the *question analysis*, the *document retrieval*, the *sentence retrieval*, and the *answer extraction* (see [3] for a detailed description of each module).

This year some minor changes were introduced regarding: (i) the handling of temporally restricted questions, (ii) the final validation of the extracted answers, and (iii) the adaptation of the system to a cross-language environment. The next subsections describe these three topics in more detail.

Additionally, other small improvements were made, such as the implementation of a priority based scheme to make question categorisation more assertive, and the use of inexact matching techniques (based on the Levenshtein distance) for partial matching of proper nouns in the sentence retrieval module. The latter was done adapting existing technology for the Portuguese spell checker⁵. Future work

⁴ SVMTool is developed by TALP Research Center NLP group, of Universitat Politècnica de Catalunya (<http://www.lsi.upc.es/~nlp/SVMTool/>).

⁵ The Portuguese spell checker is included in FLiP, together with a grammar checker, a thesaurus, a hyphenator, a verb conjugator and a translation assistant that enable different proofing level – word, sentence, paragraph and text – of European and Brazilian Portuguese. An online version is available at <http://www.flip.pt>

will address extending the use of inexact matching techniques in the document retrieval module, which will prevent the exclusion of documents where a proper noun is misspelled or spelled differently from the question, which is included in the causes of failure in the cross-language task (cf. section 4).

3.1 Handling of Temporally Restricted Questions

The QA@CLEF organisation distinguishes three types of temporal restrictions: *date* (e.g., “Who was the US president in 1962?”), *period* (e.g., “How many cars were sold in Spain between 1980 and 1995?”) and *event* (e.g., “Where did Michael Milken study before enrolling in the University of Pennsylvania?”). Our approach focuses on the two first types of restrictions. We take into account two sources of information: (i) the dates of documents, and (ii) temporal expressions in the documents' text.

The documents' dates in the collections *EFE* (for Spanish), *Público* (for European Portuguese) and *Folha de São Paulo* (for Brazilian Portuguese) are an instance of metadata. To exploit this source, we provided our system with the ability to deal with metadata information. This is a common requirement of real-world systems in most domains (the Web, digital libraries and local archives). Our procedure to deal with dates is extensible to other kinds of metadata, such as the document's title, subject, author information, etc. Restricted questions often require a hybrid search procedure, using both metadata Boolean-like queries and NLP techniques. During indexation, our system finds the adequate piece of metadata containing the date of each document which is then indexed. This makes the system able to handle Boolean-like date queries (such as “select all documents dated above 1994-10-01 and below 1994-10-15”).

A more difficult issue is the recognition of temporal expressions in NL text. While *absolute* dates like “25th April, 1974” are easy to recognise and convert to a numeric format, the same does not happen when the dates are *incomplete* (“25th April”, “April 1974”, “1974”), when the temporal expressions are less conventional (“spring 1974”, “2nd quarter of 1974”), or when we consider *temporal deictics* (like “yesterday”, “last Monday”, “20th of the current month”) that require knowledge about the discourse date.

As a first approach, we were only concerned with temporal expressions that refer to possibly incomplete absolute dates or periods. To answer temporally restricted questions, we need to perform operations over numeric representations of dates and periods, like *date comparison* (is 1974-04-25 greater, equal or lower than 1974-05-01?), *translation of time units* (add 8 days to 1974-04-25), or *checking if a date lies in a period* (is 1974-04-28 in the interval [1974-04-25, 1974-05-01]?). The need to deal with incomplete dates requires some fuzziness in these operations.

The procedure for answering a temporally restricted question is as follows: during question analysis, the temporal expression is recognised. The document retriever module *relaxes* the restriction into a larger period (starting a few days

before, and ending a few days after)⁶, and applies a metadata query to retrieve a set T_1 of those documents whose dates are relevant. Then, a regular query is applied to retrieve a set T_2 of documents containing NL temporal expressions that match the question restriction. Finally, the usual document retrieval procedure continues as in the unrestricted case, with the difference that the top-30 retrieved documents are constrained to belong to the set $T = T_1 \cup T_2$.

3.2 Answer Validation

Last year the absence of a strategy for final answer validation was one of the major causes of failure of our system. This year we have made a naïve approach to this problem. In order to strengthen the match between the question and the sentence containing the answer, we demand the following: (i) the answering sentence must match (at least partially) all the proper nouns and NEs in the question, and (ii) it must match a given amount of nominal and verbal pivots in the question. Here, a match is any correspondence of lemmas, heads of derivation, or synonyms. If a candidate answer is extracted from a sentence that fails one of these two criteria, it is discarded, although it can still be used for coherence analysis of other answers.

Notice that this approach ignores any relation among the question pivots and the extracted answer; it merely checks for matches of each pivot individually. Current work is addressing a more sophisticated strategy for answer validation through syntactic parsing. The idea is to capture the argument structure of the question, by setting a syntactical role to each pivot or group of pivots in the sentence, requiring that the answer to be extracted is contained in a specific phrase, and checking if the answering sentence actually offers enough support.

3.3 Adaptation of the System to a Cross-Language Environment

The adaptation of the system to a cross-language environment required few modifications. Our approach is self-contained, in the sense that it does not require using any external software or Web searches to perform translations. Instead, we use an ontology based direct translation, refined by means of corpora statistics.

The central piece is the multilingual ontology, also used in last year's CLEF for monolingual [3] and bilingual [8] purposes. The combination of the ontology information of all TRUST languages provides a bidirectional word/expression translation mechanism. Some language pairs are directly connected, as is the case of Portuguese and Spanish; others are connectable using the English language as an intermediate. This allows operating in a cross-language scenario for any pair of languages in the ontology (among Portuguese, Spanish, French, English, Polish and Czech⁷). For each language pair that is directly connected, translation scores are used to reflect the likelihood of each translation. These scores are computed in

⁶ This "relaxation" approach seems reasonably appropriate for newspaper corpora, since it works both for news about events that already occurred, and for announces of incoming events.

⁷ The French, Polish and Czech parts of the ontology are property respectively of Synapse Développement, TiP and University of Economics, Prague (UEP).

Table 1. Example of the Portuguese-Spanish translation for the question “Qual é o recorde mundial do salto em altura?” [What is the world record for high jump?]. Notice that the ontology includes the expression *salto em altura/salto de altura*, and that although *momento* [moment] is chosen as the preferential translation for the polysemic word *altura*, the most correct *altura* and *altitud* are also taken into account.

Pivots	Translations	Translations kept as synonyms
recorde	récord	
mundial	mundial	universal
salto em altura	salto de altura	
salto	salto	brinco, salto mortal, sobresalto, carrerilla, bronco, voltereta, cambio brusco, (...)
altura	momento	tamaño, estatura, nivel, altitud, cumbre, pináculo, grandeza, altura, instante, (...)

a background task using the Europarl parallel corpus [9], a paragraph aligned corpus containing the proceedings of the European Parliament. After a preliminary sentence alignment step, each aligned sentence is processed simultaneously for the two languages under consideration, and words from one language are associated to their translation in the second language. The number of times one word is associated to another is recorded and used to calculate the translation score. Presently, we are exploiting Europarl for other purposes, like word sense disambiguation.

In a cross-language environment, our QA system starts by instantiating independent language modules for each language (in this case, Portuguese and Spanish). Suppose that a question is asked in language X , while language Y is chosen as target. First, the question analyser performs pivot extraction and question categorisation, using the language module for X . Then, if the ontology has a direct connection from X to Y , each pivot is translated to language Y . Otherwise, each pivot is translated to English, and then the language module for Y is used to translate from English to Y . Whenever multiple translations are available, the one with the highest translation score is chosen as default, while the others are kept as synonyms (see Table 1). This is done for pivots’ lemmas, heads of derivation and synonyms. Notice that this strategy does not translate the whole question from X to Y , which would discard alternative translations of some words. At this point, the language module for Y has pivots in its own language and question categories which are language independent. So it proceeds as in the Y monolingual environment, with the document retrieval module, the sentence extractor and the answer extractor.

4 Results

To analyse the results evaluated in QA@CLEF, we divided the sets of 200 questions into five major categories regarding the type of the expected answer:

Table 2. Results by category of question

Answer →	Right				Total				Accuracy (%)			
Question ↓	PT	ES	PT	ES	PT	ES	PT	ES	PT	ES	PT	ES
			ES	PT			ES	PT			ES	PT
DEN	69	49	33	29	100	96	96	100	69.0	51.0	34.4	29.0
DEF	19	18	9	14	31	27	28	31	61.3	66.7	32.1	45.1
LOC	19	13	10	8	30	27	26	30	63.3	48.2	38.5	26.7
QUANT	11	11	9	9	18	24	24	18	61.1	45.8	37.5	50.0
TEMP	16	14	11	7	21	26	26	21	76.2	53.8	42.3	33.3
Total	134	105	72	67	200	200	200	200	67.0	52.5	36.0	33.5

denomination/designation (DEN), *definition* (DEF), *location/space* (LOC), *quantification* (QUANT) and *duration/time* (TEMP).

The general results of Table 2 show that there was a slight increase regarding last year’s Portuguese monolingual task. As for Priberam’s first participation in the Spanish monolingual task, we achieved better results than last year’s best performing system [1]. As for the cross-language environment, the accuracy is significantly lower, and relatively similar for the Portuguese-Spanish and Spanish-Portuguese tasks.

Despite the fairly satisfying results in the Spanish monolingual task, there is still a considerable gap between the performance of the Spanish and the Portuguese modules. A few reasons contributed to this. The semantic classification, the addition of proper nouns and the NEs detection rules for Spanish were just in an early stage, which led to a higher percentage of errors related to the extraction of candidate answers. The absence of a thesaurus also played a part in the document and sentence retrieval stages.

Table 3 displays the distribution of errors along the main stages of our QA system, both in the monolingual and bilingual runs.

Table 3. Reasons for wrong (W), inexact (X) and unsupported (U) answers

Question →	W+X+U				Failure (%)			
Stage ↓	PT	ES	PT	ES	PT	ES	PT	ES
			ES	PT			ES	PT
Document retrieval	21	16	28	27	31.8	16.8	21.9	20.3
Extraction of candidate answers	23	41	19	26	34.9	43.2	14.8	19.5
Choice of the final answer	15	30	36	37	22.7	31.6	28.1	27.8
NIL validation	7	8	14	12	10.6	8.4	10.9	9.0
Translation	—	—	29	28	—	—	22.7	21.1
Other	0	0	2	3	0.0	0.0	1.6	2.3
Total	66	95	128	133	100.0	100.0	100.0	100.0

In the monolingual runs, most errors occur during the extraction of candidate answers. This is related with several issues: QAPs that are badly tuned or erroneously applied due to errors in question categorisation, difficulties in writing QAPs to extract the right answer in long sentences, and a few errors due to the handling of morphological and lexical relations, as well as anaphoric relations (e.g. the ES question 80 “¿En qué año murió Bernard Montgomery?” [In what year did Bernard Montgomery die?] did not retrieve the answer “(...) su muerte, ocurrida en 1976, motivó un auténtico duelo nacional”). The inclusion of other kinds of metadata besides the date may improve the answer extraction performance, as suggested by PT question 133 “Onde ganharam os Abba o Festival da Eurovisão?” [Where did Abba win the Eurovision Song Contest?], whose answer, *Brighton*, should be inferred from the title of the news.

As for the document retrieval stage, many errors are connected with long questions with many pivots, which may have the effect of filtering out relevant documents while keeping others that contain more but less important pivots. There were also issues of matching proper nouns and NEs, which are usually the core pivots (e.g. in PT question 53 “Quem venceu a Volta a França em 1988?” [Who won the Tour of France in 1988?] the right document was not retrieved, because the sentence of the answer contained *Tour* instead of *Volta a França*); difficulties with ambiguous NEs (e.g. in PT question 189 “Quem escreveu ‘A Capital’?” [Who wrote ‘A Capital’?], ‘A Capital’ occurs frequently in the corpus as the name of a newspaper and less often as the title of a book). Furthermore, some questions could only be answered if a more sophisticated strategy of indexation was made (e.g. the answer to ES question 124 “¿Qué zar ruso murió en 1584?” lies in a document with a list of events: “EFEMERIDES DEL 17 DE MARZO (...) Defunciones (...) 1584.- Ivan ‘El Terrible’, zar ruso.”, and could only be retrieved with a different indexation strategy). Table 3 shows also that the Spanish document retrieval stage outperformed the Portuguese one, although we were expecting the opposite behaviour since the system did not use a Spanish thesaurus. The only explanation we could find is that the Portuguese questions were, on average, harder to retrieve than the Spanish ones, more of them requiring anaphora resolution and difficult matches.

Our simple strategy for answer validation led to fair results. We estimated our system's NIL recall/precision as respectively 65%/43% (PT), 60%/34% (ES), 30%/29% (PT-ES) and 30%/18% (ES-PT).

The work done for temporally restricted questions, described in the subsection 3.1, allowed the system to retrieve the right answer in 40.7% of the questions, in the Portuguese run, and 32.1%, in the Spanish one. The justification for these results is related with several different issues, mostly in the document retrieval and answer selection stages, and not necessarily with our procedure to handle this type of restriction. For instance, in PT question 155 “Contra que clube jogou o Marítimo a 2 de Abril de 1995?” [Against what team did Marítimo play on the 2nd April 1995?] the answer was not extracted because, due to tokenization, we could not retrieve the document containing the answer, *Marítimo-Tirsense*, although the

system correctly matched the date of the document within the relaxed interval set by the temporal restriction.

In the bilingual runs, one of the major causes of failure is related with translation errors, mostly with proper nouns and NEs equivalencies. For instance, in the questions “Quantos Óscares ganhou ‘A Guerra das Estrelas’?”; “O que é a Eurovisão?” and “Cuándo se suicidó Cleopatra?”, the system missed the translations *A Guerra das Estrelas/La guerra de las Galaxias*, *Eurovisão/Eurovision* and *Cleópatra/Cleopatra*. Since NEs and proper nouns are generally the most important elements for the right extraction of answers, if one fails to translate them correctly, they are bound to have a negative impact in the results. Notice that the ontology was not designed to contain proper nouns and NEs except those belonging to closed domains, like names of countries and other geographic entities. Frequently (like in the two last examples above) there are only slight spelling differences between a word and its translation. This suggests dealing with this issue by extending the use of inexact matching techniques to the document retrieval module, hence preventing the exclusion of documents by misspelling of proper nouns or NEs. This technique may also improve the monolingual usage, since in “real-world” systems it is frequent to have different spellings of proper nouns in the question and in the target documents.

5 Conclusions and Future Work

Despite the positive results, there are still a few issues that need to be solved, in order to improve the general performance of the system. We are currently enriching the Spanish lexicon with the inclusion of proper nouns, semantic features, synonyms and lexical relations.

In addition, we are implementing syntactical processing to question categorisation to capture the argument structure of the question. This will allow a more tuned answer extraction, by matching the specific roles of the question arguments in the answer. This strategy will also lead to an improvement of the cross-language performance, taking profit of the language independence of that argument structure.

Current work is being done in M-CAST project for dealing with other metadata information besides dates, like the document title, its author and subject. A task to be addressed in the future is the application of inexact matching techniques, currently used only in the sentence retrieval stage, in the document retrieval module as well. Of course, this would imply some changes in the indexation scheme. As said above, we expect this to improve not only the translation of proper nouns and NEs in cross-language environments, but also the monolingual robustness. Finally, research is being done to address word sense disambiguation. Europarl corpus [9] is being used with the aim of building a Portuguese sense disambiguated corpus, taking profit of the multilingual ontology, which enables domain specific translation, hence allowing discriminating word senses.

Acknowledgements

Priberam Informática would like to thank the partners of the NLUC consortium⁸, especially Synapse Développement, the CLEF organization and Linguatca. We would also like to acknowledge the support of the European Commission in TRUST (IST-1999-56416) and M-CAST (EDC 22249 M-CAST) projects.

References

1. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Penas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 multilingual question answering track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. Amaral, C., Laurent, D., Martins, A., Mendes, A., Pinto, C.: Design and Implementation of a Semantic Search Engine for Portuguese. In: Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004), vol. 1, pp. 247–250 (2004). Also available at <http://www.priberam.pt/docs/LREC2004.pdf>
3. Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C.: Priberam's question answering system for Portuguese. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Amaral, C., Figueira, H., Mendes, A., Mendes, P., Pinto, C.: A Workbench for Developing Natural Language Processing Tools. In: Pre-proceedings of the 1st Workshop on International Proofing Tools and Language Technologies, Patras, Greece (July 1-2, 2004). Also available at <http://www.priberam.pt/docs/WorkbenchNLP.pdf>
5. Thede, S.M., Harper, M.P.: A second-order hidden Markov model for part-of-speech tagging. In: Proceedings of the 37th Annual Meeting of the ACL, pp. 175–182. College Park, Maryland (1999)
6. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing (2nd printing). The MIT Press, Cambridge, Massachusetts (2000)
7. Giménez, J., Márquez, L.: SVMTool: A general POS tagger generator based on Support Vector Machines. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal (2004)
8. Laurent, D., Séguéla, P., Négre, S.: Cross Lingual Question Answering using QRISTAL for CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
9. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Proceedings of MT Summit X (Phuket, Thailand), pp. 79–86 (2005). Also available at: <http://www.iccs.inf.ed.ac.uk/~pkoeht/publications/europarl-mtsummit05.pdf>

⁸ See <http://www.nluc.com>.

LCC's PowerAnswer at QA@CLEF 2006

Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakor, Jonathan Clark,
and Dan Moldovan

Language Computer Corporation
Richardson, Texas 75080, USA
{mitchell, marian, moldovan}@languagecomputer.com,
<http://www.languagecomputer.com>

Abstract. This paper reports on Language Computer Corporation's first QA@CLEF participation. For this exercise, LCC integrated its open-domain PowerAnswer Question Answering system with its statistical Machine Translation engine. For 2006, LCC participated in the English-to-Spanish, French and Portuguese cross-language tasks. The approach is that of intermediate translation, only processing English within the QA system regardless of the input or source languages. The output snippets were then mapped back into the source language documents for the final output of the system and submission. What follows is a description of the system and methodology.

1 Introduction

In 2006, Language Computer's open-domain question answering system PowerAnswer [5] participated in QA@CLEF for the first time. PowerAnswer has previously participated in many other evaluations, notably NIST's TREC [1] workshop series, however, this is the first Multilingual QA evaluation the system has entered. LCC has developed its own statistical machine translation system, which was integrated with PowerAnswer for this evaluation. Since PowerAnswer is a very modular and extensible system, the integration required only a minimum of modifications for the initial approach.

The goals for this year's participation were (1) to examine how well the current QA system performs when given noisy data, such as that from automatic translation and (2) to examine the performance of the machine translation system in a question answering environment. To that end, LCC adopted an approach of intermediate translation instead of adapting the QA system to process target languages natively.

The paper presents a summary of the PowerAnswer system, the machine translation engine, the integration of the two for QA@CLEF 2006, and then follows with a discussion of results and challenges in the CLEF question topics. This year, LCC participated in the following bilingual tasks: English $\rightarrow$ French, English $\rightarrow$ Spanish, English $\rightarrow$ Portuguese.

2 Overview of LCC's PowerAnswer

Automatic question answering requires a system that has a wide range of tools available. There is no one monolithic solution for all question types or even data sources. In realization of this, LCC developed PowerAnswer as a fully-modular and distributed multi-strategy question answering system that integrates semantic relations, advanced inferencing abilities, syntactically constrained lexical chains, and temporal contexts. This section presents an outline of the system and how it was modified to meet the challenges of QA@CLEF 2006.

PowerAnswer comprises a set of strategies that are selected based on advanced question processing, and each strategy is developed to solve a specific class of questions either independently or together. A Strategy Selection module automatically analyzes the question and chooses a set of strategies with the algorithms and tools that are tailored to the class of the given question. PowerAnswer can distribute the strategies across workers in the case of multiple strategies being selected, alleviating the increase in the complexity of the question answering process by splitting the workload across machines and processors.

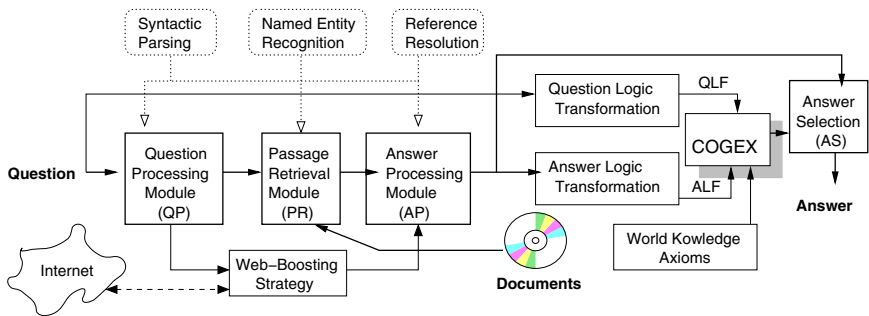


Fig. 1. PowerAnswer 2 Architecture

Each strategy is a collection of components, (1) Question Processing (*QP*), (2) Passage Retrieval (*PR*), and (3) Answer Processing (*AP*). Each of these components constitute one or more modules, which interface to a library of generic NLP tools. These NLP tools are the building blocks of the PowerAnswer 2 system that, through a well-defined set of interfaces, allow for rapid integration and testing of new tools and third-party software such as IR systems, syntactic parsers, named entity recognizers, logic provers, semantic parsers, ontologies, word sense disambiguation modules, and more. Furthermore, the components that make up each strategy can be interchanged to quickly create new strategies, if needed, they can also be distributed [10].

As illustrated in Figure 1, the role of the *QP* module is to determine (1) temporal constraints, (2) the expected answer type, (3) to process any question semantics necessary such as roles and relations, (4) to select the keywords used in retrieving relevant passages, and (5) perform any preliminary questions

as necessary for resolving question ambiguity. The *PR* module ranks passages that are retrieved by the *IR* system, while the *AP* module extracts and scores the candidate answers based on a number of syntactic and semantic features such as keyword density, count, proximity, semantic ordering, roles and entity type. All modules have access to a syntactic parser, semantic parser, a named entity recognizer and a reference resolution system through LCC's generic NLP tool libraries. To improve the answer selection, PowerAnswer takes advantage of redundancy in large corpora, specifically in this case, the Internet. As the size of a document collection grows, a question answering system is more likely to pinpoint a candidate answer that closely resembles the surface structure of the question. These features have the role of correcting the errors in answer processing that are produced by the selection of keywords, by syntactic and semantic processing and by the absence of pragmatic information. Usually, the final decision for selecting answers is based on logical proofs from the inference engine COGEX [7]. For this year's QA@CLEF, however, the logic prover was disabled in order to better evaluate the individual components of the QA architecture. COGEX was evaluated on multilingual data in CLEF's Answer Validation Exercise [13], where it was the top performer in both Spanish and English.

3 Overview of Translation Engine

The translation system used at LCC implements phrase-based statistical machine translation [2], the core translation engine is the open-source Phramer [12] system, developed by one of LCC's engineers. Phramer in turn implements and extends the phrase-based machine translation algorithms implemented by Pharaoh [4]. A more detailed description of the MT solution adopted for Multilingual QA@CLEF can be found in [11]. The translation system is trained using the Europarl Corpus [3], which provides between 600,000 and 800,000 pairs of sentences (sentences in English paired with the translation in another European language). LCC followed the training procedure described in the Pharaoh training manual¹ to generate the phrase table required for translation.

In order to translate entire documents, the core translation engine is augmented with (1) tokenization, (2) capitalization, and (3) de-tokenization. The tokenization process is performed on the original documents (in French, Portuguese or Spanish), in order to convert the sentences to space-separated entities, in which the punctuation and the words are isolated. The step is required because the statistical machine translation core engine accepts only lowercased tokenized input. The capitalization process follows the translation process and it restores the casing of the words. The capitalization tool uses three-gram statistics extracted from 150 million words from the English GigaWord 2nd Edition corpus, augmented with two heuristics: (i) First word will always be uppercased; (ii) If the words appear also in the foreign documents, the casing is preserved (this rule is very effective for proper nouns and named entities).

¹ <http://www.iccs.inf.ed.ac.uk/~pkoeht/training.tgz>

4 PowerAnswer-Phramer Integration

LCC's cross-language solution for Question Answering is based on automatic translation of the documents in the source language (English). QA is performed on a collection consisting only of English documents. The answers were converted back into the target language (the original language of the documents) by aligning the translation with the original document (finding the original phrase in the original document that generated the answer in English); when this method failed, the system falls back to machine translation (source $\rightarrow$ target).

4.1 Passage Retrieval

Making use of PowerAnswer's modular design, LCC developed three different retrieval methods, settling on the first of these for the final experiment.

1. use an index of English words, created from the translated documents
2. use an index of foreign words (French, Spanish or Portuguese), created from the original documents
3. use an index of English words, created from the original documents in correlation with the translation table

The first solution is the default solution. The entire target language document collection is translated into English, processed through the set of NLP tools and indexed for querying. Its major disadvantage is the computational effort required to translate the entire collection. It also requires updating the English version of the collection when one improves the quality of the translation. Its major advantage is that there are no additional costs during question answering (the documents are already translated). This passage retrieval method is illustrated in Figure 2. The second solution, as seen in Figure 3, requires minimum effort dur-

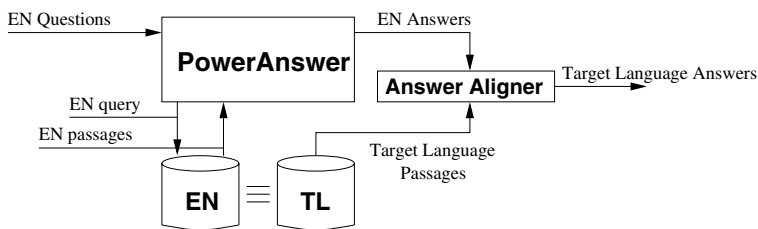


Fig. 2. Passage Retrieval on English documents (default)

ing indexing (the document collection is indexed in its native language). In order to retrieve the relevant documents, the system translates the keywords of the IR query (the query submitted by PowerAnswer to the Lucene-based ² IR system) with alternations as the new IR query (step 1). The translation of keywords

² <http://lucene.apache.org/>

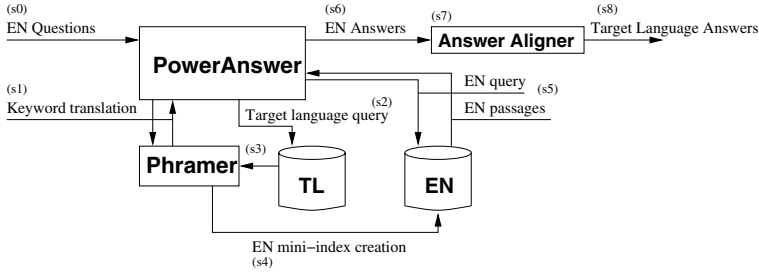


Fig. 3. Passage Retrieval on Target Language documents

is performed using Phramer, by generating n-best translations. This translated query is submitted to the target language index (step 2). The documents retrieved by this query are then dynamically translated into English using Phramer (step 3). The system uses a cache to store translated documents so that IR query reformulations and other questions that might retrieve the same documents will not need to be translated again. The set of translated documents is indexed into a mini-collection (step 4) and the mini-collection is re-queried using the original English-based IR query (step 5). For example, the boolean IR query in English (“poem” AND “love” AND “1922”) is translated into French as (“poeme” AND (“aiment” OR “aimer” OR “aimez” OR “amour”) AND “1922”) with the alternations. This new query will return 85 French documents. Some of them do not contain “love” in their automatic translation (but the original document contains “aiment”, “aimer”, “aimez” or “amour”). By re-querying the translated sub-collection (that contains only the translation of those 85 documents) the system retrieves only 72 English documents to be passed to PowerAnswer.

The advantage of the second method is that minimum effort is required during collection preparation. Also, the collection preparation might not be under the control of the QA system (i.e. it can be web-based). Also, improvements in the MT engine can be reflected immediately in the output of the integrated system. The disadvantage is that more computation is required at run-time for translating the IR query and the documents dynamically.

The third alternative extracts during indexing the English words that might be part of the translation and indexes the collection accordingly. The process doesn’t involve lexical choice - all choices are considered possible. The set of keywords is determined using the translation table, and collects all words that are part of the translation lattice ([4]). Determining only the words according to the translation table (semi-translation) is approximately 10 times faster than the full translation. The index is queried using the original IR query generated by PowerAnswer (with English keywords). After the initial retrieval, the algorithm is similar to the second method: translate the retrieved documents, re-query the mini-collection. The advantage is the much smaller indexing time when compared with the first method, besides all the advantages of the second method. Also, it has all the disadvantages of the second method, except that it doesn’t

require IR query translation. Because preliminary testing proved that there aren't significant differences in recall between the three methods and because the first method is fastest after the document collection is prepared, only the first method was used for the final evaluation.

4.2 Answer Processing

For each of the above methods, PowerAnswer returns the exact answer and the supporting source sentence (all in English). These answers are aligned to the corresponding text in the target language documents. The final output of the system is the response list in the target language with the appropriate supporting snippets. If the alignment method fails, the English answers are converted directly into the target language as the final responses.

5 Results

The integrated multilingual PowerAnswer system was tested on 190 English $\rightarrow$ Spanish, 190 English $\rightarrow$ French and 188 English $\rightarrow$ Portuguese factoid and definition questions and 10 English $\rightarrow$ Spanish, 10 English $\rightarrow$ French and 12 English $\rightarrow$ Portuguese list questions. For QA@CLEF, the main score is the overall accuracy, the average of SCORE(q), where SCORE(q) is defined for factoids and definition questions as 1 if the top answer for q is assessed as correct, 0 otherwise. Also included are the Mean Reciprocal Rank (MRR) and the Confidence Weighted Score (CWS) that judges how well a system returns correct answers higher in the ranked list of answers. Table 1 illustrates the final results of Language Computer's efforts in its first participation at QA@CLEF for 2006.

6 Error Analysis and Challenges in 2006

There were several sources of errors in LCC's submission, including one major source that accounts for a 50% loss in accuracy. The sources of errors include: translation misalignments, tokenization errors, and data processing errors.

6.1 Translation Misalignments

Because the version of PowerAnswer used this year is *monolingual*, the system design for *multilingual* question answering involved translating documents dynamically for processing through the QA system and mapping the responses back into the source language documents. This resulted in several opportunities for error. While the translation of the documents into English did introduce noise into the data such as mistranslations, words that were not translated and should have been or words that should not have been translated and were, it did not affect the QA system as much as expected. By far the greatest source of errors was the alignment between the English answers and the source documents which

Table 1. LCC’s QA@CLEF 2006 Factoid/Definition Results

Source	Accuracy	CWS	MRR
Spanish	20.00%	0.04916	0.2000
French	21.05%	0.04856	0.2623
Portuguese	8.51%	0.01494	0.1328

Table 2. LCC’s Factoid/Definition Results in English

Source	Position 1 Acc.	Top 5 Acc.	Submission Acc.
Spanish	40.00%	63.16%	20.00%
French	40.53%	64.74%	21.05%
Portuguese	38.83%	62.23%	8.51%

produced the final response list. This source accounts for roughly a 50% loss of accuracy for the tasks from PowerAnswer output to final submission. Table 2 compares the PowerAnswer English accuracy versus the mapped submission accuracy. For Portuguese, there was also a submission format error that accounts for the great difference in accuracy when compared to the Spanish and French results. This table also demonstrates the potential this system has once these misalignment errors are corrected.

6.2 Data Errors

Questions that contained common nouns in the source language question were one example of data processing errors. Question 83 (EN-PT) is *Who is the director of the film “Caro diario”?*. Here, the noun “Caro” was translated “expensive” when being indexed in English. Hence, the query keyword “Caro” as it is in the question could not be found in the translated collection by the IR system, causing the passage to receive a lowered score when retrieved on less keywords.

Sometimes the wording of the questions also affected the system in a negative way. There were some spelling changes that were unrecoverable by the system, such “Huan Karlos” for *Juan Carlos* (EN-PT #5). There were also a few instances of keywords for which PowerAnswer was unable to generate the correct alternation, for example “celebrated” in EN-ES #75 *In which year was the Football World Cup celebrated in the United States?*, where the correct alternation would have been a synonym for “to host”.

The scoring of definition questions tends to be subjective, so there were cases where LCC believes an answer returned warranted at worst an inexact but was judged wrong, such as EN-PT #84 *What is the Unhcr?*. PowerAnswer responded with “O Acnur (Alto Comissariado das Nações Unidas para Refugiados) consultou o Brasil e mais 30 países sobre a possibilidade de acolher um grupo de 5.000 refugiados da ex-Iugoslávia”, which was judged as wrong. Additionally, the definition question strategy often returned answer snippets that were a full sentence and were judged as inexact because of their length. One example is EN-PT

question 34 *Who was Alexander Graham Bell?*. PowerAnswer returns the full sentence containing the important nugget “*A empresa foi fundada em 1885 e entre os sócios estava Alexander Graham Bell, o inventor do telefone*”, where the final part “inventor of the telephone” was all that was necessary.

7 Conclusions

While this year's performance was not what was hoped due to some major errors in the final stages of processing answers, there is great potential for better results in the following years. One large step forward is to make the core of PowerAnswer more language-independent. The English dependency comes from the NLP tools more than the modules of PowerAnswer, where the language dependence occurs primarily in Question Processing. LCC's multilingual QA capabilities will see great improvement though developing a set of multilingual NLP tools, including syntactic parsers and semantic relation extractors, abstracting out the English-dependent components of PowerAnswer and building language-independent as well as some language-specific question processing. This work will be eased by the flexible design of PowerAnswer and the company's proprietary libraries of generic NLP tool interfaces.

References

1. Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A.: Employing Two Question Answering Systems in TREC-2005. In: Text REtrieval Conference (2005)
2. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of HLT/NAACL 2003, Edmonton, Canada (2003)
3. Koehn, P.: Europarl: A Parallel Corpus for SMT. MT Summit 2005 (2005)
4. Koehn, P.: Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In: Frederking, R.E., Taylor, K.B. (eds.) AMTA 2004. LNCS (LNAI), vol. 3265, Springer, Heidelberg (2004)
5. Moldovan, D., Harabagiu, S., Clark, C., Bowden, M.: PowerAnswer 2: Experiments and Analysis over TREC 2004. In: Text REtrieval Conference (2004)
6. Moldovan, D., Clark, C., Harabagiu, S.: Temporal Context Representation and Reasoning. In: Proceedings of IJCAI, Edinburgh, Scotland (2005)
7. Moldovan, D., Clark, C., Harabagiu, S., Maiorano, S.: COGEX A Logic Prover for Question Answering. In: Proceedings of the HLT/NAACL (2003)
8. Moldovan, D., Novischi, A.: Lexical chains for Question Answering. In: Proceedings of COLING, Taipei, Taiwan (August 2002)
9. Moldovan, D., Rus, V.: Logic Form Transformation of WordNet and its Applicability to Question Answering. In: Proceedings of ACL, France (2001)
10. Moldovan, D., Srikanth, M., Fowler, A., Mohammed, A., Jean, E.: Synergist: Tools for Intelligence Analysis. In: NIMD Conference, Arlington, VA (2006)
11. Olteanu, M., Suriyentrakorn, P., Moldovan, D.: Language Models and Reranking for Machine Translation. In: NAACL 2006 Workshop On SMT (2006)
12. Olteanu, M., Davis, C., Volosen, I., Moldovan, D.: Phramer - An Open Source Statistical Phrase-Based Translator. In: NAACL 2006 Workshop On SMT (2006)
13. Tatu, M., Iles, B., Moldovan, D.: Automatic Answer Validation using COGEX. In: Cross-Language Evaluation Forum (CLEF) (2006)

Using Syntactic Knowledge for QA^{*}

Gosse Bouma, Ismail Fahmi, Jori Mur, Gertjan van Noord,
Lonneke van der Plas, and Jörg Tiedemann

Information Science, University of Groningen,
PO Box 716 9700 AS Groningen The Netherlands
{g.bouma,i.fahmi,j.mur,g.j.m.van.noord,m.l.e.van.der.plas,
j.tiedemann}@rug.nl

Abstract. We describe the system of the University of Groningen for the monolingual Dutch and multilingual English to Dutch QA tasks. First, we give a brief outline of the architecture of our QA-system, which makes heavy use of syntactic information. Next, we describe the modules that were improved or developed especially for the CLEF tasks, among others incorporation of syntactic knowledge in IR, incorporation of lexical equivalences and coreference resolution, and a baseline multilingual (English to Dutch) QA system, which uses a combination of Systran and Wikipedia (for term recognition and translation) for question translation. For non-list questions, 31% (20%) of the highest ranked answers returned by the monolingual (multilingual) system were correct.

1 Introduction

Joost (see figure 1) is a question answering system for Dutch which performs full syntactic analysis of the question and all text in the document collection. Answers are extracted by pattern matching over dependency relations, and potential answers are ranked, among others, by computing the syntactic similarity between the question and the sentence from which the answer is extracted. Apart from the three classical components *question analysis*, *passage retrieval* and *answer extraction*, the system also contains a component (called *qatar*) for extracting answers off-line. All components in our system rely heavily on syntactic analysis, which is provided by Alpino [3], a wide-coverage dependency parser for Dutch.

Question analysis produces a set of dependency relations (i.e. the result of syntactic analysis), it determines the question type, and identifies keywords (for IR). Depending on the question type the next stage is either passage retrieval or table look-up. If the question type matches one of the table categories, this table will be searched for an answer. Tables are created off-line for facts that

^{*} This research was carried out as part of the research program for *Interactive Multimedia Information Extraction*, IMIX, financed by NWO, the Dutch Organisation for Scientific Research.

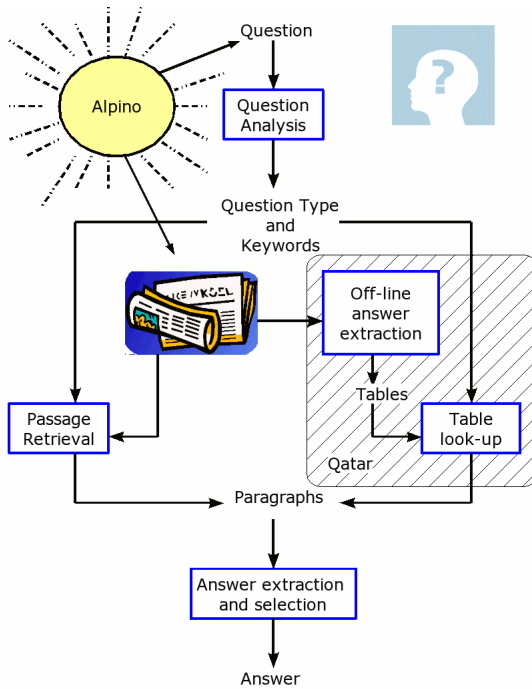


Fig. 1. System architecture of Joost

frequently occur in fixed patterns. We store these facts as potential answers together with the IDs of the paragraphs in which they were found. If table look-up cannot be used, we follow the other path through the QA-system to the passage retrieval component. The IR-engine selects relevant paragraphs to be passed on to subsequent modules for extracting the actual answer.

The final processing stage is answer extraction and selection. The input to this component is a set of paragraph IDs, either provided by Qatar or by the IR system. For questions that are answered by means of table look-up, the tables provide an exact answer string. In this case the paragraph context is used only for ranking the answers. For other questions, answer strings have to be extracted from the paragraphs returned by IR. Answer extraction patterns are similar to those used for off-line extraction, but typically more general (and more noisy). Answers are ranked on the basis of frequency of the answer, overlap in named entities between question and answer, syntactic similarity between question and answer, and (estimated) reliability of the extraction pattern used. Finally, the highest ranked answers are returned to the user. More detailed descriptions of the system can be found in [1] and [2].

In the rest of the paper, we focus on components of the system that were revised or developed for CLEF 2006, and on discussion of the results.

2 Linguistically Informed Information Retrieval

The information retrieval component identifies relevant paragraphs in the corpus to narrow down the search space for subsequent modules. Answer containing paragraphs that are missed by IR are lost for the entire system. Hence, IR performance in terms of recall is essential. Furthermore, high precision is also desirable as IR scores are used for ranking potential answers.

Given a full syntactic analysis of the CLEF text collection, it becomes feasible to exploit linguistic information as a knowledge source for IR. Using Apache's IR system Lucene [4], we can index the document collection along various linguistic dimensions, such as part of speech tags, named entity classes, and dependency relations. We defined several *layers* of linguistic features and feature combinations and included them as index fields. In our current system we use the following layers: text (stemmed text tokens), root (root forms), RootPos (root forms concatenated with wordclass labels), RootRel (root forms concatenated with the name of the dependency relation to their head words), RootHead (dependent-head bigrams using root forms), RootRelHead (dependent-head bigrams with the type of relation between them), compound (compounds identified by Alpino), ne (named entities), neLOC, nePER, neORG (only location, person, or organisation names, respectively), and neTypes (labels of named entities in the paragraph). The layers are filled with appropriate data extracted from the analysed corpus.

Each of the index fields defined above can be accessed using Lucene's query language. Complex queries combining keywords for several layers can be constructed. Queries to be used in our system are constructed from the syntactically analysed question. We extract linguistic features in the same way as done for building the index. The task now is to use this rich information appropriately. The selection of keywords is not straightforward. Keywords that are too specific might harm the retrieval performance. It is important to carefully select features and feature combinations to actually improve the results compared to standard plain text retrieval.

For the selection and weighting of keywords we applied a genetic algorithm trained on previously collected question answer pairs. For constructing a query we defined further keyword restrictions to make an even more fine-grained selection. For example, we can select RootHead keywords from the question which have been tagged as nouns. Each of these (possibly restricted) keyword selections can be weighted with a numeric value according to their importance for retrieval. They can also be marked as "required" using the '+' character in Lucene's query syntax. All keyword selections are then concatenated in a disjunctive way to form the final query. Figure 2 gives an example.

Details of the genetic optimisation process are given in [5]. As the result of the optimisation we obtain an improvement of about 19% over the baseline using standard plain text retrieval (i.e. the text layer only) on unseen evaluation data. It should be noted that this improvement is not solely an effect of using root forms or named entity labels, but that many of the features that are assigned a high weight by the genetic algorithm refer to layers that make use of dependency information.

```

text:(stelde Verenigde Naties +embargo +Irak)
ne:(Verenigde_Naties^2 Verenigde^2 Naties^2 Irak^2)
RootHead:(Irak/tegen embargo/stel_in)
neTypes:(YEAR)

```

Fig. 2. An example IR query from the question *Wanneer stelde de Verenigde Naties een embargo in tegen Irak ?* (When did the United Nations declare the embargo against Iraq?) using plain text tokens, named entities with boost factor 2, RootHead bigrams for nouns, and the named entity class for the question type.

3 Coreference Resolution for Off-Line Question Answering

Off-line answer extraction has proven to be very effective and precise. The main problem with this technique is the lack of coverage. One way to increase the coverage is to apply coreference resolution. For instance, the age of a person may be extracted from snippets such as:

- (1) a. de 26-jarige Steffi Graf (*the 26-year old Steffi Graf*)
- b. Steffi Graf....de 26-jarige tennisster (*Steffi Graf...the 26-year old tennis player*)
- c. Steffi Graf....Ze is 26 jaar. (*Steffi Graf...She is 26 years old*)

If no coreference resolution is applied, only patterns in which a named entity is present, such as (1-a) will match. Using coreference resolution, we can also extract the age of a person from snippets such as (1-b) and (1-c), where the named entity is present in a preceding sentence.

We selected 12 answer types that we expect to benefit from coreference resolution: **age**, **date/location of birth**, **age/date/location/cause of death**, **capital**, **inhabitants**, **founder**, **function**, and **winner**. Applying the basic patterns to extract facts for these categories we extracted 64,627 fact types. We adjusted the basic patterns by replacing the slot for the named entity with a slot for a pronoun or a definite NP.

Our strategy for resolving definite NPs is based on knowledge about the categories of named entities, so-called instances (or categorised named entities). Examples are *Van Gogh* IS-A *painter*, *Seles* IS-A *tennis player*. We acquired instances by scanning the corpus for apposition relations and predicate complement relations. We scan the left context of the definite NP for named entities from right to left. For each named entity we encounter, we check whether it occurs together with the definite NP as a pair on the instance list. If so, the named entity is selected as the antecedent of the NP. As long as no suitable named entity is found we select the next named entity and so on until we reach the beginning of the document. If no named entity is found that forms an instance pair with the definite NP, we select simply the first preceding named entity.

We applied a similar technique for resolving pronouns. Again we scan the left context of the anaphor (now a pronoun) for named entities from right to left.

We implemented a preference for proper nouns in the subject position. For each named entity we encounter, we check whether it has the correct NE-tag and number. If we are looking for a person name, we do another check to see if the gender is correct.¹ After having resolved the anaphor, the corresponding fact containing the antecedent named entity was added to the appropriate table.

We estimated the number of additional fact types we found using the estimated precision scores (on 200 manually evaluated facts). Coreference resolution extracted approximately 39,208 (60.7%) additional facts. 5.6% of these involve pronouns, and 55.2% definite NPs. The number of facts we extracted by the pronoun patterns is quite low. We did a corpus investigation on a subset of the corpus which consisted of sentences containing terms relevant to the 12 selected question types². In only 10% of the sentences one or more pronouns appeared. This outcome indicates that the possibilities of increasing coverage by pronoun resolution are inherently limited.

4 Lexical Equivalences

One of the features that is used to rank potential answers to a question is the amount of syntactic similarity between the question and the sentence from which the answer is taken. Syntactic similarity is computed as the proportion of dependency relations from the question which have a match in the dependency relations of the answer sentence. In [6], we showed that taking syntactic equivalences into account (such as the fact that a *by*-phrase in a passive is equivalent to the subject in the active, etc.) makes the syntactic similarity score more effective.

In the current system, we also take lexical equivalences into account. That is, given two dependency relations $\langle \text{Head}, \text{Rel}, \text{Dependent} \rangle$ and $\langle \text{Head}', \text{Rel}, \text{Dependent}' \rangle$, we assume that they are equivalent if both **Head** and **Head'** and **Dependent** and **Dependent'** are near-synonyms. Two roots are considered near synonyms if they are identical, synonyms, or spelling variants, or if one is an abbreviation, genitive form, adjectival form, or compound suffix of the other.

A list of synonyms (containing 118K root forms in total) was constructed by merging information from EuroWordNet, dictionary websites, and various encyclopedias (which often provide alternative terms for a given lemma keyword). The spelling of person and geographical names entities tends to be subject to a fair amount of variation and the spelling used in a question is not necessarily the same as the one used in a paragraph which provides the answer. Using edit distance to detect spelling variation tends to be very noisy. To improve the precision of this method, we restricted ourselves to person names, and imposed the additional constraint that the two names must occur with the same function in our database of functions (used for off-line question answering). Thus, *Felipe Gonzalez* and *Felippe Gonzales* are considered to be variants only if they are

¹ We created a list of boy's names and girl's names by downloading such lists from the Internet. To be accepted as the correct antecedent, the proper name should not occur on the name list of the opposite sex of the pronoun.

² Terms such as "geboren" (*born*), "stierf" (*died*), "hoofdstad" (*capital*) etc.

known to have the same function (e.g. prime-minister of Spain). Currently, we recognize 4500 pairs of spelling variants. The compound rule applies when one of the words is a compound suffix of the other. It also cover multi word terms analyzed as a single word by the parser (i.e. *colitis ulcerosa*).

We tested the effect of incorporating lexical equivalences on questions from previous CLEF tasks. Although approximately 8% of the questions receives a different answer when lexical equivalences are incorporated, the effect on the overall score is negligible. We suspect that this is due to the fact that in the definition of synonyms, no distinction is made between various senses of a word, and the equivalences defined for compounds tend to introduce a fair amount of noise (e.g. the *Calypso-queen* of the Netherlands is not the same as the *queen* of the Netherlands). It should also be noted that most lexical equivalences are not taken into consideration by the IR-component. This probably means that some relevant documents (especially those containing spelling variants of proper names) are missed.

5 Definition Questions

Definition questions can ask either for a definition of a named entity (*What is Lusa?*) or a concept (*What is a cincinatto*). We used appositions (*the Portuguese press agency Lusa*), nominal modifiers (*milk sugar (saccharum lactis)*), or (*ofwel*) disjunctions with *ofwel* (the Dutch variant of *or* which is often used to indicate alternative names for a concept (*melksuiker ofwel saccharum lactis*, *milk sugar or saccharum lactis*), predicative complements (*milk sugar is (called/known as) saccharum lactis*), and predicative modifiers (*composers such as Joonas Kookonen*) to find potential answers. As some of these patterns tend to be very noisy, we also check whether there exists an ISA-relation between the head noun of the definition, and the term to be defined. ISA-relations are collected from named entity – noun appositions (48K) and head noun – concept pairs (136K) extracted from definition sentences in an automatically parsed version of Dutch Wikipedia. Definition sentences were identified automatically (see [7]). Answers for which a corresponding ISA-relation exists in Wikipedia are given a higher score.

For the 40 definition questions in the Dutch QA test set, 18 received a correct first answer (45%), which is considerably better than the overall performance on non-list questions (31%). We consider 7 of the 40 definition questions to be concept definition questions. Of those, only 1 was answered correct.

6 Temporally Restricted Questions

Sometimes, questions contain an explicit date:

- (2) a. Which Russian Tsar died in 1584?
- b. Who was the chancellor of Germany from 1974 to 1982?

To provide the correct answer to such questions, it must be ensured that there is no conflict between the date mentioned in the question and temporal information present in the text from which the answer was extracted.

If a sentence contains an explicit date expression, this is used as *answer date*. A sentence is considered to contain an explicit date if it contains a temporal expression referring to a date (*2nd of August, 1991*) or a relative date (*last year*). The denotation of the latter type of expression is computed relative to the date of the newspaper article from which the sentence is taken. Sentences which do not contain an explicit date are assigned an *answer date* which corresponds to the date of the newspaper from which the sentence is extracted.

For questions which contain an explicit date, this date is used as the *question date*. For all other questions, the *question date* is nil. The *date score* of a potential answer is 0 if the *question date* is nil, 1 if answer and question date match, and -1 otherwise.

There are 31 questions in the Dutch QA test set which contain an explicit date, and which we consider to be temporally restricted questions. Our monolingual QA system returned 11 correct first answers for these questions (10 of correctly answered questions ask explicitly for a fact from 1994 or 1995). The performance of the system on temporally restricted questions is similar to the performance achieved for (non-list) questions in general (31%).

7 Multilingual QA

We have developed a baseline English to Dutch QA-system which is based on two freely available resources: Systran and Wikipedia. For development, we used the CLEF 2004 multieight corpus. [8]

The English source questions are converted into an HTML file, which is translated automatically into Dutch by Systran.³ These translations are used as input for the monolingual QA-system described above.⁴

This scenario has a number of obvious drawbacks: (1) translations often result in grammatically incorrect sentences, (2) even if a translation can be analyzed syntactically, it may contain words or phrases that were not anticipated by the question analysis module, and (3) named entities and (multiword) terms are not recognized. We did not spend any time on fixing the first and second potential problem. While testing the system, it seemed that the parser was relatively robust against grammatical irregularities. We did notice that question analysis could be improved, so as to take into account peculiarities of the translated questions.

³ Actually, we used the Babelfish interface to Systran, <http://babelfish.altavista.digital.com/>

⁴ For English to Dutch, the only alternative on-line translation service seems to be Freetranslation (www.freetranslation.com). When testing the system on questions from the multieight corpus, the results from Systran seemed slightly better, so we decided to use Systran only.

The third problem seemed most serious to us. It seems Systran fails to recognize many named entities and multiword terms. The result is that these are translated on a word by word basis, which typically leads to errors that are almost certainly fatal for any component (starting with IR) which takes the translated string as starting point. To improve on the treatment of named entities and terms, we extracted from English Wikipedia all pairs of lemma titles and their cross-links to the corresponding link in Dutch Wikipedia. Terms in the English input which are found in the Wikipedia list are escaped from automatic translation and replaced by their Dutch counterparts directly. This potentially avoids three types of errors: (1) the term should not be translated, but it is by Systran (*Jan Tinbergen* is translated as *Januari Tinbergen*), (2) the term is not translated by Systran, but it should (*Pippi Longstocking* is not translated, but should be translated as *Pippi Langkous*), (3) the term should be translated, but it is translated wrongly by Systran (*Pacific Ocean* is translated as *Vreedzame Oceaan*, but should be translated as *Stille Oceaan* or *Grote Oceaan*).

Of the 200 input questions, 48 contained terms that matched an entry in the bilingual term database extracted from Wikipedia. 4 of the marked terms are incorrect (*Martin Luther* instead of *Martin Luther King* is marked as a term, *nuclear power* instead of *nuclear power plants* is marked as a term, *prime-minister* is translated as *minister-voorzitter* rather than as *minister-president* or *premier*, and *the game* is incorrectly recognized as a term (it matches the name of a movie in Wikipedia) and not translated).

Although the precision of recognizing terms is high, it should be noted that recall could be much better. Unrecognized terms are translated on a word by word basis (*Olympic Winter Games* is translated as *Olympische Spelen van de Winter* instead of *Olympische Winterspelen*, the name *Jack Soden* (which should not be translated) is translated as *Hefboom Soden*). In addition, unrecognized proper names may show up as discontinuous strings in the translation (i.e. *What did Yogi Bear steal* → *Wat Yogi stal de Beer*).

Although the performance of the multilingual system is a good deal less than that of the monolingual system, there actually are a few questions which are answered correctly by the bilingual system, but not by the monolingual system. This is due to the fact that the (more or less) automatic word by word translations in these cases match more easily with the answer sentences than the corresponding Dutch questions (which are paraphrases of the English sentence).

8 Evaluation and Error Analysis

The results from the CLEF evaluation are given in figure 3.

The monolingual system assigned only 13 questions a question type for which a table with potential answers was extracted off-line. For only 5 of those, an answer is found off-line. This suggests that the effect of off-line techniques on the overall result is relatively small. As off-line answer extraction tends to be

Q type	#	# correct	% correct	MRR
Factoid Questions	146	40	27.4	
Definition Questions	40	18	45	
Temporally Restricted ⁶	1	0	0	
Non-list questions	187	58	31	0.346
List Questions	13	15/65 answers correct (P@5 = 0.23)		

Q type	#	# correct	% correct	MRR
Factoid Questions	147	27	18.4	
Definition Questions	39	11	28.2	
Temporally Restricted	1	0	0	
Non-list questions	187	38	20.3	0.223
List Questions	13	4/37 answers correct (P@5 = 0.06)		

Fig. 3. Official CLEF scores for the monolingual Dutch task (top) and bilingual English to Dutch task (bottom)

more accurate than IR-based answer extraction, it may also explain why the results for the CLEF 2006 task are relatively modest.⁵

If we look at the scores per question type for the most frequent question types (as they were assigned by the question analysis component), we see that definition questions are answered relatively well (18 out of 40 of the first answers correct), that the scores for general WH-questions and location questions are in line with the overall score (16 out of 52 and 8 out of 25 correct), but that measure and date questions are answered poorly (3 out of 20 and 3 out of 15 correct). On the development-set (of 800 questions from previous CLEF tasks), all of these question types perform considerably better (the worst scoring question type are measure questions, which still finds a correct first answer in 44% of the cases).

A few questions are not answered correctly because the question type was unexpected. This is true in particular for the (3) questions of the type *When did Gottlob Frege live?*.

Somewhat suprisingly, question analysis also appears to have been an important source of errors. We estimate that 23 questions were assigned a question type that was either wrong or dubious. Dubious assignments arise when question analysis assigns a general question type (i.e. *person*) where a more specific question type was available (i.e. *founder*).

Attachment errors of the parser are the source of small number of mistakes. For instance, Joost replies that O.J. Simpson was accused of *murder on his ex-wife*, where this should have been *murder on his ex-wife and a friend*. As the conjunction is misparsed, the system fails to find this constituent. Different attachments also cause problems for the question *Who was the German chancellor*

⁵ For development, we used almost 800 questions from previous CLEF tasks. For those questions, almost 30% of the questions are answered by answers that were found off-line. 75% of the first answers for those questions is correct. Overall, the development system finds almost 60% correct first answers.

between 1974 and 1982?. It has an almost verbatim answer in the corpus (*the social-democrat Helmut Schmidt, chancellor between 1974 and 1982*), but since the temporal restriction is attached to the verb in the question, and the noun *social-democrat* in the answer, this answer is not found.

The performance loss between the bilingual and the monolingual system is approximately 33%. This is somewhat more than the differences between multilingual and monolingual QA reported for many other systems (see [9] for an overview). However, we do believe that it demonstrates that the syntactic analysis module is relatively robust against the grammatical anomalies present in automatically translated input. It should be noted, however, that 19 out of 200 questions cannot be assigned a question type, whereas this is the case for only 4 questions in the monolingual system. Adapting the question analysis module to typical output produced by automatic translation, and improvement of the term recognition module (by incorporating a named entity recognizer and/or more term lists) seems relatively straightforward, and might lead to somewhat better results.

References

1. Bouma, G., Mur, J., van Noord, G., van der Plas, L., Tiedemann, J.: Question answering for Dutch using dependency relations. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedeman, J.: Linguistic knowledge and question answering. *Traitement Automatique des Langues*
3. Bouma, G., van Noord, G., Malouf, R.: Alpino: Wide-coverage computational analysis of Dutch. In: *Computational Linguistics in The Netherlands 2000*, Rodopi, Amsterdam (2001)
4. Jakarta, A.: Apache Lucene - a high-performance, full-featured text search engine library (2004), <http://lucene.apache.org/java/docs/index.html>
5. Tiedemann, J.: Improving passage retrieval in question answering using NLP. In: *Proceedings of the 12th Portuguese Conference on Artificial Intelligence (EPIA)*, Covilhã, Portugal. LNAI Series, Springer, Heidelberg (2005)
6. Bouma, G., Mur, J., van Noord, G.: Reasoning over dependency relations for QA. In: Benamara, F., Saint-Dizier, P. (eds.) *Proceedings of the IJCAI workshop on Knowledge and Reasoning for Answering Questions (KRAQ)*, Edinburgh, pp. 15–21 (2005)
7. Fahmi, I., Bouma, G.: Learning to identify definitions using syntactic features. In: Basili, R., Moschitti, A. (eds.) *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*, Trento, Italy (2006)
8. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Penas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the clef 2004 multilingual question answering track. In: Peters, C., Clough, P.D., Jones, G.J.F., Gonzalo, J., Kluck, M., Magnini, B. (eds.) *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*. LNCS, vol. 3491, Springer, Heidelberg (2005)
9. Ligozat, A.L., Grau, B., Robba, I., Vilat, A.: Evaluation and improvement of cross-lingual question answering strategies. In: Peñas, A., Sutcliffe, R. (eds.) *EACL workshop on Multilingual Question Answering*, Trento, Italy (2006)

A Cross-Lingual German-English Framework for Open-Domain Question Answering

Bogdan Sacaleanu and Günter Neumann

LT-Lab, DFKI, Saarbrücken, Germany
{bogdan, neumann}@dfki.de

Abstract. The paper describes QUANTICO, a cross-language open domain question answering system for German and English. The main features of the system are: use of preemptive off-line document annotation with syntactic information like chunk structures, apposition constructions and abbreviation-extension pairs for the passage retrieval; use of online translation services, language models and alignment methods for the cross-language scenarios; use of redundancy as an indicator of good answer candidates; selection of the best answers based on distance metrics defined over graph representations. Based on the question type two different strategies of answer extraction are triggered: for factoid questions answers are extracted from best IR-matched passages and selected by their redundancy and distance to the question keywords; for definition questions answers are considered to be the most redundant normalized linguistic structures with explanatory role (i.e., appositions, abbreviation's extensions). The results of evaluating the system's performance by CLEF were as follows: for the best German-German run we achieved an overall accuracy (ACC) of 42.33% and a mean reciprocal rank (MRR) of 0.45; for the best English-German run 32.98% (ACC) and 0.35 (MRR); for the German-English run 17.89% (ACC) and 0.17 (MRR).

1 Introduction

QUANTICO is a cross-language open domain question answering system developed for both English and German factoid and definition question. It uses a common framework for both monolingual and cross-language scenarios, with different workflow settings for each task and different configurations for each type of question. For tasks with different languages on each end of the information flow (questions and documents) we cross the language barrier rather on the question than on the document side by using free online translation services, linguistic knowledge and alignment methods. An important aspect of QUANTICO is the triggering of specific answering strategies by means of control information that has been determined by the question analysis tool, e.g. question type and expected answer type, see [3] for more details. Through the offline annotation of the document collection with several layers of linguistic information (chunks, appositions, named entities, sentence boundaries) and their use in the retrieval process, more accurate and reliable information units are being considered for answer extraction, which is based on the assumption that redundancy is a good indicator of information suitability. The answer selection component normalizes

and represents the context of an answer candidate as a graph and computes its appropriateness in terms of the distance between the answer and question keywords.

We will begin giving a short overview of the system and presenting its working for both factoid and definition questions in monolingual and cross-language scenarios. We will then continue with a short description of each component and close the paper with the presentation of the CLEF evaluation results and the error analysis outcome.

2 System Overview

QUANTICO uses a common framework for both monolingual and cross-language scenarios, but with different configurations for each type of question (definition or factoid) and different workflow settings for each task (DE2DE, DE2EN or EN2DE).

Concerning the workflow settings, the following things are to be mentioned. For the monolingual scenario (DE2DE) the workflow is as follows (according to the architecture in the Figure 1): 1-4-5-6/7 with the last selection depending on the question type. For a cross-language scenario, the workflow depends on the language of the question: for German questions and English documents (DE2EN) the workflow is

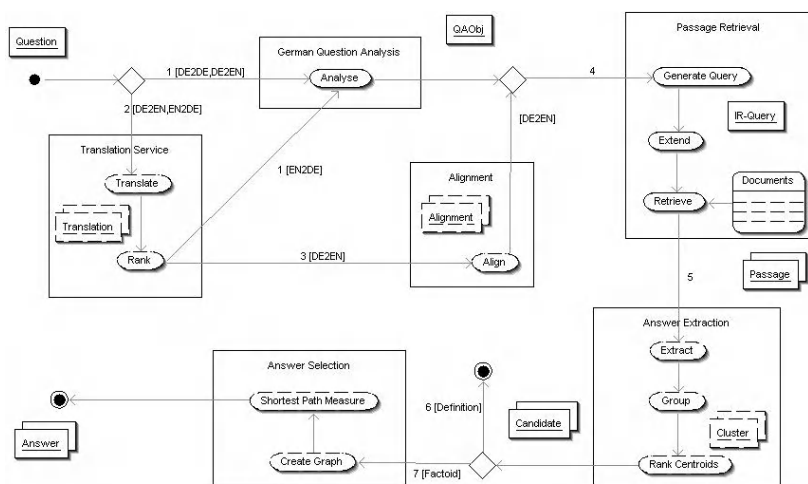


Fig. 1. System Architecture

1-2-3-4-5-6/7, that is, the question is first analyzed, then translated and aligned to its translations, so that based on the generated *QAObj* and the alignments a new English *QAObj* is being computed; for English questions and German documents (EN2DE) the workflow is 2-1-4-5-6/7, that is, the question is first translated and then the best translation – determined according to linguistic completeness – is being analyzed resulting in a *QAObj*. The difference in the system's workflow for the cross-language scenario comes with our choice of analyzing only German questions, since our analysis component, based on the SMES parser [1], is very robust and accurate. In

the presence of a *Question Analysis* component with similar properties for English questions, the workflow would be the same (1-2-3-4-5-6/7) independent of the question's language.

Regarding the component configurations for each type of question (definition or factoid) the difference is to be noted only in the *Passage Retrieval* and *Answer Extraction* components. While the *Retrieve* process for the factoid questions builds on classic Information Retrieval methods, for definition questions is merely a look-up procedure in a repository of offline extracted syntactic structures as appositions, chunks and abbreviation-extension pairs. For the *Answer Extraction* component the distinction consists in different methods of computing the clusters of candidate answers: for factoid question, where the candidates are usually named entities or chunks, is based on co-reference (*John ~ John Doe*) and stop-word removal (*of death ~ death*), while for definition questions, where candidates can vary from chunks to whole sentences, is based on topic similarity (*Italian designer ~ the designer of a new clothes collection*).

3 Component Description

Following is a description of QUANTICO's individual components along with some examples.

3.1 Question Analysis

In the context of a QA system or information search in general, we interpret the result of a NL question analysis as a *declarative description of search strategy and control information*, see [3]. Consider, for example, the NL question result for the question “*In welcher Stadt fanden 2002 die olympischen Winterspiele statt?*” (*The Olympic winter games took place 2002 in which town?*), where the value of the tag *a-type* represents the expected answer type, *q-type* the answer control strategy, and *q-focus* and *q-scope* additional constraints for the search space:

```
<QOBJ msg="quest" id="qId0" lang="de" score="1">
  <NL-STRING id="qId0">
    <SOURCE id="qId0" lang="de">In welcher Stadt fanden 2002 die
      olympischen Winterspiele statt?</SOURCE>
    </NL-STRING>
  <QA-control>
    <Q-FOCUS>Stadt</Q-FOCUS>
    <Q-SCOPE>stattfind_winter#spiel</Q-SCOPE>
    <Q-TYPE restriction="TEMP">C-COMPLETION</Q-TYPE>
    <A-TYPE type="atomic">LOCATION</A-TYPE>
  </QA-control>
  <KEYWORDS>
    <KEYWORD id="kw0" type="UNIQUE">
      <TK pos="V" stem="statt#find">fanden</TK>
    </KEYWORD>
    <KEYWORD id="kw1" type="UNIQUE">
```

```

<TK pos="N" stem="stadt">Stadt</TK>
</KEYWORD>
<KEYWORD id="kw2" type="UNIQUE">
<TK pos="NUMERAL" stem="2002">2002</TK>
</KEYWORD>
<KEYWORD id="kw3" type="UNIQUE">
<TK pos="A" stem="olympisch">olympischen</TK>
</KEYWORD>
<KEYWORD id="kw4" type="UNIQUE">
<TK pos="N" stem="winter#spiel">Winterspiele</TK>
</KEYWORD>
</KEYWORDS>
<EXPANDED-KEYWORDS />
<NE-LIST>
<NE id="ne0" type="DATE">2002</NE>
</NE-LIST>
</QOBJ>

```

Parts of the information can already be determined on basis of local lexico-syntactic criteria (e.g., for the Wh-phrase *where* we can simply infer that the expected answer type is *location*). However, in most cases we have to consider larger syntactic units in combination with the information extracted from external knowledge sources. For example for a definition question like “*What is a battery?*” we have to combine the syntactic and type information from the verb and the relevant NP (e.g., combine definite/indefinite NPs together with certain auxiliary verb forms) in order to distinguish it from a description question like “*What is the name of the German Chancellor?*” In our QAS, we are doing this by following a two-step parsing schema:

- in a first step a full syntactic analysis is performed using the robust parser SMES and
- in a second step a question-specific semantic analysis.

During the second step, the values for the question tags *a-type*, *q-type*, *q-focus* and *q-scope* are determined on basis of syntactic constraints applied on the dependency analysis of relevant NP and VP phrases (e.g., considering agreement and functional roles), and by taking into account information from two small knowledge bases. They basically perform a mapping from linguistic entities to values of the questions tags, e.g., trigger phrases like *name_of*, *type_of*, *abbreviation_of* or a mapping from lexical elements to expected answer types, like *town*, *person*, and *president*. For German, we additionally perform a *soft retrieval match* to the knowledge bases taking into account on-line compound analysis and string-similarity tests. For example, assuming the lexical mapping *Stadt* → *LOCATION* for the lexeme *town*, then automatically we will also map the nominal compounds *Hauptstadt* (capital) and *Großstadt* (large city) to *LOCATION*.

A main aspect in the adaptation and extension of the question analysis component for the Clef-2006 task concerned the recognition of the question type, i.e., simple factoid and list factoid questions, definition questions and the different types of the temporally restricted questions. Because of its high degree of modularity of the question analysis component, the extension only concerns the semantic analysis sub-component. Here, additional syntactic-semantic mapping constraints have been

implemented that enriched the coverage of the question grammar, where we used the question set of the previous Clef campaigns as our development set.

3.2 Translation Services and Alignment

We have used two different methods for responding questions asked in a language different from the one of the answer-bearing documents. Both employ online translation services (Altavista, FreeTranslation, etc.) for crossing the language barrier, but at different processing steps, i.e. before and after formalizing the user information need into a *QAObj*.

The *a priori-method* translates the question string in an earlier step, resulting in several automatic translated strings, of which the best one is analyzed by the *Question Analysis* component and passed on to the *Passage Retrieval* component. This is the strategy we use in an English–German cross-lingual setting. To be more precise: the English source question is translated into several alternative German questions using online MT services. Each German question is then parsed with SMES, our German parser. The resulting query object is then weighted according to its linguistic well-formedness and its completeness with respect to the query information (question type, question focus, answer-type).

The assumption behind this weighting scheme is that “a translated string is of greater utility for subsequent processes than another one, if its linguistic analysis is more complete or appropriate.”

The *a posteriori-method* translates the formalized result of the *Query Analysis* component by using the question translations, a language modeling tool and a word alignment tool for creating a mapping of the formal information need from the source language into the target language. We illustrate this strategy in a German–English setting along two lines (using the following German question as example: “*In welchem Jahrzehnt investierten japanische Autohersteller sehr stark?*”):

- translations as returned by the on-line MT systems are being ranked according to a language model
 - *In which decade did Japanese automakers invest very strongly?* (0.7)
 - *In which decade did Japanese car manufacturers invest very strongly?* (0.8)
- translations with a satisfactory degree of resemblance to a natural language utterance (i.e. linguistically well-formedness), given by a threshold on the language model ranking, are aligned based on several filters: dictionary filter - based on MRD (machine readable dictionaries), PoS filter - based on statistical part-of-speech taggers, and cognates filter - based on string similarity measures (dice coefficient and LCSR (lowest common substring ratio)).

In: [in:1.0] 1.0
welchem: [which:0.5] 0.5
Jahrzehnt: [decade:1.0] 1.0
investierten: [invest:1.0] 1.0
japanische: [Japanese:0.5] 0.5
Autohersteller: [car manufacturers:0.8, auto makers:0.1] 0.8
sehr: [very:1.0] 1.0
stark: [strongly:0.5] 0.5

3.3 Passage Retrieval

The preemptive offline document annotation refers to the process of annotating the document collections with information that might be valuable during the retrieval process by increasing the accuracy of the hit list. Since the expected answer type for factoid questions is usually a named entity type, annotating the documents with named entities provides for an additional indexation unit that might help to scale down the range of retrieved passages only to those containing the searched answer type. The same practice applies for definition questions given the known fact that some structural linguistic patterns (appositions, abbreviation-extension pairs) are used with explanatory and descriptive purpose. Extracting these kinds of patterns in advance and looking up the definition term among them might return more accurate results than those of a search engine.

The *Generate Query* process mediates between the question analysis result *QObj* (answer type, focus, keywords) and the search engine (factoid questions) or the repository of syntactic structures (definition questions) serving the retrieval component with information units (passages). The *Generate Query* process builds on an abstract description of the processing method for every type of question to accordingly generate the *IRQuery* to make use of the advanced indexation units. For example given the question “*What is the capital of Germany?*”, since named entities were annotated during the offline annotation and used as indexing units, the *Query Generator* adapts the *IRQuery* so as to restrict the search only to those passages having at least two locations: one as the possible answer (*Berlin*) and the other as the question’s keyword (*Germany*), as the following example shows:

+text:capital+text:Germany+neTypes:LOCATION +LOCATION:2.

It is often the case that the question has a semantic similarity with the passages containing the answer, but no lexical overlap. For example, for a question like “*Who is the French prime-minister?*”, passages containing “*prime-minister X of France*”, “*prime-minister X ... the Frenchman*” and “*the French leader of the government*” might be relevant for extracting the right answer. The *Extend* process accounts for bridging this gap at the lexical level, either through look-up of unambiguous resources or as a side-effect of the translation and alignment process (see [4]).

In the context of the participation to CLEF two different settings have been considered for the retrieval of relevant passages for factoid questions: one in which a passage consists of only a sentence as retrieval unit from a document, and a second one with a window of three adjoining sentences for a passage. Concerning the query generation, only keywords with following part-of-speeches have been considered for retrieval: nouns, adjective and verbs, whereby only nouns are mandatory to occur in the matching relevant passages. In case of empty hit list retrieval, the query undergoes a relaxation process maintaining only the focus of the question and the expected answer type (as computed by the *Analyse* component) as mandatory items:

<i>Question:</i>	Which country did Joe visit?
<i>IR-Query:</i>	+neTypes:LOCATION +text:country^4 +text:Joe
text:visit	
<i>Relaxed IR-Query:</i>	+neTypes:LOCATION +text:country^4 text:Joe
text:visit	

3.4 Answer Extraction

The *Answer Extraction* component is based on the assumption that the redundancy of information is a good indicator for its suitability. The different configurations of this component for factoid and definition questions reflect the distinction of the answers being extracted for these two question types: simple chunks (i.e. named entities and basic noun phrases) and complex structures (from phrases through sentences) and their normalization. Based on the control information supplied by the *Analyse* component (*q-type*), different extraction strategies are being triggered (noun phrases, named entities, definitions) and even refined according to the *a-type* (definition as sentence in case of an OBJECT, definition as complex noun phrase in case of a PERSON).

Whereas the *Extract* process for definition questions is straightforward for cases in which the offline annotation repository lookup was successful, in other cases it implies an online extraction of those passage-units only that might bear a resemblance to a definition. The extraction of these passages is attained by matching them against a lexico-syntactic pattern of the form:

<Searched Concept> <definition verb> .+

whereby *<definition verb>* is being defined as a closed list of verbs like “is”, “means”, “signify”, “stand for” and so on.

For factoid questions having named entities or simple noun phrases as expected answer type the *Group* (normalization) process consists in resolving cases of coreference, while for definition questions with complex phrases and sentences as possible answers more advanced methods are being involved. The current procedure for clustering definitions consists in finding out the focus of the explanatory sentence or the head of the considered phrase. Each cluster gets a weight assigned based solely on its size (definition questions) or using additional information like the average of the IR-scores and the document distribution for each of its members (factoid questions).

3.5 Answer Selection

Using the most representative sample (centroid) of the answer candidates’ best-weighted clusters, the *Answer Selection* component sorts out a list of top answers based on a distance metric defined over graph representations of the answer’s context. The context is first normalized by removing all functional words and then represented as a graph structure. The score of an answer is defined in terms of its distance to the question concepts occurring in its context and the distance among these.

In the context of the participation to CLEF a threshold of five best-weighted clusters has been chosen and all their instances, not only their centroids, have been considered for a thorough selection of the best candidate.

4 Evaluation Results

We participated in three tasks: DE2DE (German to German), EN2DE (English to German) and DE2EN (German to English), with two runs submitted for each of the first two tasks. The second runs submitted (*dfki062*) were distinct in that the context of the retrieved passages was consisting of three sentences compared to the other runs (*dfki061*) with only one sentence per passage. A detailed description of the achieved results can be seen in Table 1.

Compared to the results from last year [3], we were able to keep our performance for the monolingual German task DE2DE (2005 edition: 43.50%). For the task English to German we were able to improve our result (2005 edition: 25.50%) and for the task German to English we observed a decrease (2005 edition: 23.50%).

Table 1. System Performance - Details

Run ID	Right		W	X	U	F	D	T	P@NL	NIL [20]		
	#	%	#	#	#	%	%	%	%	F	P	R
<i>dfki061dede_M</i>	80	42.32	95	6	8	38.81	56.75	29.54	25.93	0.35	0.28	0.45
<i>dfki062dede_M</i>	63	33.33	114	4	8	30.92	43.24	22.72	33.33	0.32	0.27	0.4
<i>dfki061ende_C</i>	62	32.8	117	3	6	28.94	48.64	22.72	10	0.31	0.21	0.6
<i>dfki062ende_C</i>	50	26.45	130	5	3	22.36	43.24	20.45	10	0.33	0.22	0.65
<i>dfki061deen_C</i>	34	17.89	147	9	0	17.33	20	22.5	20	0.25	0.17	0.44

Table 2 resumes the distribution of the right, inexact and unsupported answers over the first three ranked positions as delivered by our system, as well as the accuracy and MRR for each of the runs. The figures refer only to the runs submitted for the DE2DE and EN2DE tasks, since the DE2EN task evaluated just the first answer presented by the systems.

Two things can be concluded from the answer distribution of Table 2: first, there are a fair number of inexact and unsupported answers that show performance could be

Table 2. Distribution of Answers

Run ID	# Right			# inexact			# Unsupported			Accuracy	MRR
	1st	2 nd	3rd	1st	2nd	3rd	1st	2nd	3rd		
<i>dfki061dede_M</i>	80	8	7	6	6	1	8	4	1	42.32	45.67
<i>dfki062dede_M</i>	63	15	3	4	5	3	8	0	2	33.33	37.83
<i>dfki061ende_C</i>	62	5	7	3	4	2	6	3	0	32.8	35.36
<i>dfki062ende_C</i>	50	10	2	5	4	2	3	2	1	26.45	29.45
<i>dfki061deen_C</i>	34	-	-	9	-	-	0	-	-	17.89	17.89

improved with a better answer extraction; second, the fair number of right answers among the second and third ranked positions indicate that there is still place for improvements with a more focused answer selection.

5 Error Analysis

Since the actual edition of the question-answer Gold Standard has not yet been released at the time of writing this paper, the error analysis was performed only to the runs submitted for the tasks having German as target language – for which we had access to the question-answer pairs.

The results of the analysis can be grouped along two lines: conceptual and functional. The functional errors relate to the following components used:

- named entity annotation,
- on-line translation services,

while the conceptual ones refer to decisions and assumptions we made during the development of the system:

- answer and supporting evidence are to be found within a sentence
- Answer selection for instances of top five clusters might suffice
- questions and answer contexts share a fair amount of lexical items
- definition extraction strategies are exclusive

Following we will shortly explain the above-mentioned issues and provide some examples for clarity where needed.

Functional – Named Entity

The named entity tool used (LingPipe [5]), being a statistical based entity extractor, has a very good coverage and precision on annotating the document collection, where lots of context data are available, but its performance drops when using the same model for annotating short questions. Since our Query Generator component builds on using named entities as mandatory items to restrain the amount of relevant passages retrieved, failure to consistently annotate entities on both sides (question and document) results in most cases in unusable units of information and therefore wrong answers.

Functional – Translation Services

Failure of correctly translating the question from a source language to the target language can have critical results when the information being erred on represents the focus or belongs to the scope of the question. Following are several examples of miss-translations that resulted in incorrect IR-queries generation and therefore wrong answers.

“Lord of the Rings” → “Lord der Ringe” vs. “Herr der Ringe”
 “states” → “Zustände, Staate” vs. „Bundesländer“
 „high“ → „hoh, stark“ vs. „hoch“
 „Pointer Stick“ → „Zeigerstock“ vs. „Pointer Stick“
 „Mt.“ (Mount) → „Millitorr“ vs. „Mt.“

Conceptual – Answer and Supporting Evidence within a Sentence

Considering a sentence as the primary information and retrieval unit together with using the named entities as index tokens and querying terms, produced very good results in case of relatively short factoid questions where the answer and the supporting evidence (as question keywords) are to be found within the same sentence. Nevertheless, a fair amount of longer questions can only be answered by either looking at immediately adjoining sentences or using anaphora and co-reference resolution methods between noun phrases. Although LingPipe has a named entity co-reference module, it does not cover non-NE cases, which account for correctly answering some questions.

Conceptual – Answer Selection on Top Five Clusters

Looking to cover the scenario described in the previous issue, a run using three adjacent sentences as retrieval unit has been evaluated (*dfki062*). Correctly identifying answers to most of the questions by assuming scattered supporting evidence over adjoining sentences, this method invalidated some of the correctly answered factoid questions in the previous setting. The reason for that was that increasing the size of the retrieval unit produced more clusters of possible candidates and in several cases the clusters containing the correct answer were not ranked among top five and were not considered for a final selection.

Conceptual – Lexical Items Sharing between Question and Answer Context

The assumption that the question and the context of the correct answer share a fair amount of lexical items is being reflected both in the IR-query generation, although the *Expand* component might lessen it, and the answer selection. This assumption impedes the selection of correct answers that have a high semantic but little lexical overlap with the question. Some examples of semantic related concepts with no lexical overlap are as follow:

birthplace $\diamond$ born
homeland $\diamond$ born
monarch $\diamond$ king
profession $\diamond$ designer

Conceptual – Definition Extraction Strategies are Exclusive

Four extraction strategies are employed to find the best correct answer for definition questions: looking for abbreviation-extension pairs, extracting named entities with their appositions, looking for the immediately left-adjoining noun phrase and extracting definitions according to some lexico-syntactic patterns. Although the methods are quite accurate, there are cases in which either they extract false positives or the definition is inexact. Since all four strategies are exclusive, when one of them has been triggered it returns on finding a possible definition without giving a chance to any other strategy to complete. Because of this competing nature of the actual implementation, some wrong or inexact definitions are preferred over more accurate explanations.

6 Conclusions

We have presented a framework for both monolingual and cross-lingual question answering for German/English factoid and definition questions. Based on a thorough analysis of the question, different strategies are considered and alternative work-flows and components are triggered depending on the question type. Through the preemptive off-line annotation we placed some domain knowledge (i.e. named entities, appositions, abbreviations) on the document collection, so that a more effective passage retrieval approach can be used.

Intuitive assumptions regarding the unit of retrieval granularity (i.e. at sentence level) and the overlap of lexical information between the question and the relevant units have lead to promising results in the CLEF evaluation campaign, though the error analysis revealed some cases for which these premises do not hold. These are the entry points for further research to be pursued, both at the functional and conceptual level.

Acknowledgments

The work presented in this paper has been funded by the BMBF projects HyLaP (FKZ: 01 IW F02) and SmartWeb (FKZ: 01 IMD 01A).

References

- 1 Neumann, G., Piskorski, J.: A shallow text processing core engine. *Computational Intelligence* 18(3), 451–476 (2002)
- 2 Neumann, G., Sacaleanu, B.: Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering System. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004*. LNCS, vol. 3491, pp. 411–422. Springer, Heidelberg (2005)
- 3 Neumann, G., Sacaleanu, B.: Experiments on Cross-Linguality and Question-Type Driven Strategy Selection for Open-Domain QA. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 429–438. Springer, Heidelberg (2006)
- 4 Sacaleanu, B., Neumann, G.: Cross-Cutting Aspects of Cross-Language Question Answering Systems. In: *Proceedings of the EACL workshop on Multilingual Question Answering - MLQA'06*, Trento, Italy (2006)
- 5 Alias-i.LingPipe1.7 (2006), <http://www.alias-i.com/lingpipe>

Cross Lingual Question Answering Using QRISTAL for CLEF 2006

Dominique Laurent, Patrick Séguéla, and Sophie Nègre

Synapse Développement

33 rue Maynard,

31000 Toulouse, France

{dlaurent, p.seguela, sophie.negre}@synapse-fr.com

Abstract. QRISTAL [9] is a question answering system making intensive use of natural language processing both for indexing documents and extracting answers. It ranked first in the EQueR evaluation campaign (Evalda, Technolanguge [3]) and in CLEF 2005 for monolingual task (French-French) and multilingual task (English-French and Portuguese-French). This article describes the improvements of the system since last year. Then, it presents our benchmarked results for the CLEF 2006 campaign and a critical description of the system. Since Synapse Développement is participating to Quaero project, QRISTAL is most likely to be integrated in a mass market search engine in the forthcoming years.

1 Introduction

QRISTAL (French acronym for "Question Answering Integrating Natural Language Processing Techniques") is a cross lingual question answering system for French, English, Italian, Portuguese, Polish and Czech. It was designed to extract answers both from documents stored on a hard disk and from Web pages by using traditional search engines (Google, MSN, AOL, etc.). Qristal is currently used in the M-CAST European project of E-content (22249, Multilingual Content Aggregation System based on TRUST Search Engine). Anyone can assess the Qristal technology for French at www.qristal.fr. Note that the testing corpus for the testing web page is the grammar handbook proposed at <http://www.synapse-fr.com>.

For each language, a linguistic module analyzes questions and searches for potential answers. For CLEF 2006, the French, English and Portuguese modules were used for question analysis. Only the French module was used for answers extraction. The linguistic modules were developed by different companies. They share however a common architecture and similar resources (general taxonomy, typology of questions and answers and terminological fields).

For French, our system is based on the Cordial technology. It massively uses NLP tools, such as syntactic analysis, semantic disambiguation, anaphora resolution, metaphor detection, handling of converses, named entities extraction as well as conceptual and domain recognition. As the product is being marketed, the linguistic resources need to be permanently updated and it required a constant optimization of the various modules so that the software remains extremely fast. Users are now accustomed to obtain something that looks like an answer within a very short time, not exceeding **two** seconds.

2 Architecture

The architecture of the Qristal system is described in different articles (see [1], [2], [8], [9], [10], [11], [12]). Qristal is a complete engine for indexation and answers extraction. However, it doesn't index the Web. Indexing is processed only for documents based on disks. Web search uses a meta-search engine we have implemented. As we will see in the conclusion, our participation to Quaero project should change this way of use by tagging semantically the Web pages.

Our company is responsible for the indexing process of Qristal. Moreover, it ensures the integration and interoperability between all linguistic modules. Both English and Italian modules were developed by Expert System Company. The Portuguese module was developed by the Priberam Company which also takes part in CLEF 2005 for Portuguese monolingual and in CLEF 2006 for Spanish and Portuguese monolingual, and for Spanish-Portuguese and Portuguese-Spanish multilingual tasks. The Polish module was developed by the TIP Company. The Czech module is developed by the University of Economics of Prague (UEP). These modules were developed within the European projects TRUST [8] (Text Retrieval Using Semantic Technologies) and M-CAST (Multilingual Content Aggregation System based on TRUST Search Engine).

2.1 Multicriteria Indexing

While indexing documents, the technology automatically identifies the document language of and the system calls the corresponding language module. There are as many indexes as languages identified in the corpus. Documents are treated per blocks. The size of each block is approximately 1 kilobyte. Block limits are settled on the end of sentences or paragraphs. This size of block (1 kb) appeared to be optimal during our tests. Some indexes relate to blocks like fields or taxonomy whereas other relate to words, like idioms or named entities.

Each linguistic module processes a syntactic and semantic analysis for each block to be indexed. It fills a complete structure of data for each sentence. This structure is passed to the general processor that uses it to increment the various indexes. This description is accurate for the French module. Other language modules are very close to that framework but don't always include all its elements. For example, English and Italian modules do not include an indexing based on heads of derivation.

Texts are converted into Unicode. Then, they are divided into one kilobyte blocks. This reduces the index size as only the number of occurrences per block is stored for a given lemma. This number of occurrences is used to infer the relevance of each block while searching a given lemma in the index. In fact we here use lemmas but the system stores heads of derivation and not lemmas. For example, *symmetric*, *symmetrical*, *asymmetry*, *dissymmetrical* or *symmetrize* will be indexed in the same entry : *symmetry*.

Each text block is analyzed syntactically and semantically. Considering results of this analysis, 8 different indexes are built for:

- heads of derivation. A head of derivation can be a sense for a word. In French, the verb *voler* has 2 different meanings (*to steal* or *to fly*). The meaning "dérober" (*to steal*) will lead to *vol* (*robbery*), *voleur* (*thief*) or *voleuse* (*female thief*). The second meaning, "se mouvoir dans l'air" (*to fly*), will lead

to *vol* (flight), *volant* (flying as an adjective), *voleter* (to flutter) or *envol* (taking flight) and all its forms.

- proper names. If they appear in our dictionaries.
- idioms. Those idioms are listed in our idioms dictionaries. They encompass approximately 50 000 entries, like *word processing*, *fly blind* or *as good as your word*.
- named entities. Named entities are extracted from texts. *George W. Bush* or *Defense Advanced Research Project Agency* are named entities.
- concepts. Concepts are nodes of our general taxonomy. 2 levels of concepts are indexed. The first level lists 256 categories, like "visibility". The second level, actually the leaves of our taxonomy, lists 3387 subcategories, like "lighting" or "transparency",
- fields. 186 fields, like "aeronautics", "agriculture", etc.,
- question and answer types for categories like "distance", "speed", "definition", "causality", etc.,
- keywords of the text.

For each language, the indexing process is similar. Extracted data are the same. Thus, the handling of those data is independent of their original language. This is particularly important for cross language question answering.

For the French language, the rate of correct grammatical disambiguation (distinction between name-verb-adjective-adverb) is higher than 99%. The rate of semantic disambiguation is approximately 90% for 9 000 polysemous words and approximately 30 000 senses for these words. Note that this number of senses is markedly inferior to the Larousse one (Larousse is one of the most famous French dictionaries). Note however that our idioms dictionary covers a large number of the senses mentioned in this kind of dictionaries. The indexing speed varies between 200 and 400 Mo per hour with a Pentium 3 GHz, according to the size and number of indexed files.

Indexing question types is undoubtedly one of the most original aspects of our system. While the analysis of the blocks is being made, possible answers are located. For example, a name of function for a person (like *baker*, *minister*, *director of public prosecutions*), a date of birth (like *born on April 28, 1958*), a causality (like *due to snow drift* or *because of freezing*), a consequence (like *leading to serious disruption* or *facilitating the management of the traffic*). This caused the block to be indexed like being able to provide an answer for a given question type.

Currently, our question typology includes 86 types of questions. Those types are divided into two subcategories: factual types and non factual types. Factual types are dimension, surface, weight, speed, percentage, temperature, price, number of inhabitants or work of art. Nonfactual types are form, possession, judgement, goal, causality, opinion, comparison or classification. For CLEF 2006, results were as follows:

	French	English 1	English 2	Portuguese
Good choice	96.5 %	93.5 %	83.0 %	91.0 %

Fig. 1. Success rate for question type analysis

These rates are very close to CLEF 2005 results, because we only improved the French module and elaborated a new English module. The "English 1" corresponds to the Synapse English module and the "English 2" corresponds to the Expert Systems English module.

Building a keyword index for each text is also peculiar to our system. Dividing text into blocks made it compulsory. Isolated blocks cannot explicitly mention main subjects of the original text although sentences of these blocks relate to these subjects. The keyword index makes it possible to add contextual information about the main subjects of the text for blocks. Keywords can be a concept, a person, an event, etc.

2.2 Answer Extraction

After the user has keyed in his/her question, it is syntactically and semantically analyzed by the system. Question type is inferred. We would like here to draw the attention to the fact that questions are shorter than texts. This lack of context makes the semantic analysis of the question more dubious. That's why the semantic analysis processed on the question is more comprehensive than the analysis processed on texts. Moreover, users have the possibility to interactively force a sense. This possibility, however, was not used for CLEF as the entire process was automatic.

The result of the semantic analysis of the question is a weight for each sense of each word recognized as a pivot. For example, sense 1 is recognized with 20%, sense 2 with 65% and sense 3 with 15%. This weight, together with synonyms, question and answer types or concepts, is considered while searching the index. Thus all senses of a word are taken into account during the index search. This prevents from dramatic consequences due to errors in the semantic disambiguation while making the most of good analysis.

After question analysis, all indexes are searched and the best ranked blocks are analyzed again. As one can notice on figure 2, the analysis of the selected blocks is close to the analysis processed while indexing or question analyzing. On top of this "classic" analysis, a weight for each sentence is inferred. This weight is based on the number of words, synonyms and named entities found in this sentence, the presence of an answer corresponding to the question type and a correspondence between the fields and domain.

After this analysis, sentences are ranked. Then, an additional analysis is processed to extract named entities, idioms or lists that match the answer. This extraction relies on the syntactic characteristics of those groups.

For a question on a corpus located on a hard disk, the response time is approximately 1,3 seconds with a Pentium 3 GHz. On the Web, first answers are provided after 2 seconds. Then the system computes a progressive refining during ten seconds, according to user's parameters like the number of words, the number of analyzed pages, etc.

We tested many answer justification modules, mostly implemented from Web [4], [7] or [15]. Our technology enables, as an option, to use such a module of justification. It consists in searching the web with the words of the question looking for potential answers the system inferred. However this process is seldom selected by users as it increases the response time of a few seconds. It was not used in CLEF 2006 either. The only justification module we used was an internal module which makes the most of the semantic information for proper names enclosed in our dictionaries. For more than 40 000 proper names, we possess information about the country of origin, the

year of birth and death, the function for people, country, the area and population for a city, etc. We think this justification module is at the origin of some "unjustified" answers. As a matter of fact, it caused the system to rank first a text including the answer even if the system did not find any clear justification of that answer in the text.

For cross language question answering, English is used as pivot language. The fact that most users are only interested in documents in their own language and English motivated that choice. Thus, for cross language answering, the system processes generally only one translation. For this evaluation, both Portuguese to French and Italian to French runs required two translations: from source language to English and then from English to French. QRISTAL does not use any Web Services for translation because of response time. Only words or idioms recognized as pivots are translated.

3 Improvements Since CLEF 2005

For CLEF 2006, we used our same technology and system, in mono and multilingual mode[9], but with a few improvements, such as :

The ontology has been revised all through the year, mainly to emend errors or categorisation. To maintain compatibility with other language ontology, no new category has been added nor deleted.

The dictionaries have been updated, in particular proper names and expressions. This updating effort, while being continuous, has been increased last year, but not the extent to be ready, with other resources, for the said evaluation. This being the case for the dictionary of nominal expressions, leaping from 55 000 expressions to over 100 000 and not integrated in the assessed Qristal version.

A multilingual (French,English,Spanish,Italian,Portuguese) lexicon of translated proper names has been implemented. It includes more than 5 000 proper names and acronyms, mainly toponyms, (countries, provinces, towns) but also name of people, in particular names in Arabic, Russian, Chinese, which spelling differs according to the languages. This lexicon has played an important role in the improvement of the CLEF 2006 results over CLEF 2005, for the Portuguese-French pair, as the English translation was avoided.

The English-French and French-English dictionaries have also been revised and increased with now more than 200 000 translations of words or expressions. The impact of this improvement has been measured in comparing the CLEF 2005 results with the former dictionaries versus the CLEF 2006 results using the improved ones. Only one question in Portuguese to French and two questions in English to French find an additional answer with the new dictionaries.

3.1 Improvement of the Algorithms

Our syntactic analyser has not noticeably been improved. However, according to the non-yet final results of another benchmark evaluation, our analyser is given as the most performing and , above all, robust for the French language. It appears that it is still yet very far from the complete detection of the complex syntactic structures. So, for the Subject-Verb relation, it detects exactly the subject and the verb in only 9 cases out of 10. But this must be moderated by the fact , the evaluation is carried out on all types of corpora, including emails and chats, somehow more difficult to analyse.

The module « search of the category of the question » has been improved but this impacts only on the monolingual French-French part. For the M-CAST European project, our engine was enriched of numerous utilities to manage very large volumes of data and to satisfy “client-php server” structures, but these have no impact on the performances.

3.2 New English Module

The main difference between the engine assessed at CLEF 2005 and CLEF 2006, is due to the in-house development of a new English language processing module. On the basis of our ancient syntactic analyser and english-based linguistic resources not yet completed, we used a beta version of our new English module. It carries out the syntactic and semantic analysis of the question, determines the type of the question, the pivot-words and synonyms, then transfers the results to the French module that implements the requested translations of the same, to finish with the use of the language independent data (type of the question, categories of the ontology, etc.).

4 Results for CLEF 2006

QRISTAL was evaluated for CLEF 2006 for French to French, English to French (Synapse Module and Expert Systems module), and Portuguese to French. That is 1 monolingual and 3 multilingual campaigns. For each one of these tasks, we processed only one run. Note that results obtained in CLEF 2006 could have been obtained with our June 2006 commercial version of our Qristal software.

For French to French and for the pairs English-French and Portuguese-French, these results are better than those we obtained for the CLEF 2005. For French-French, that is 68% for CLEF 2006 and 64% for EQueR 2005. For English-French, that is 44,5 % for CLEF 2006 and 39,5% for CLEF 2005, and for Portuguese-French, that is 47% for CLEF 2006 and 36,5 % for CLEF 2005.

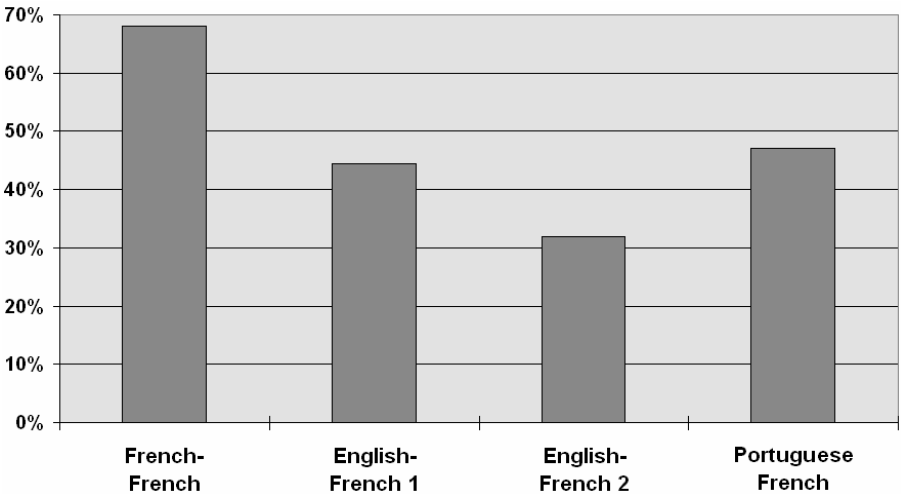


Fig. 2. Results of the general task

Results per category are as follows:

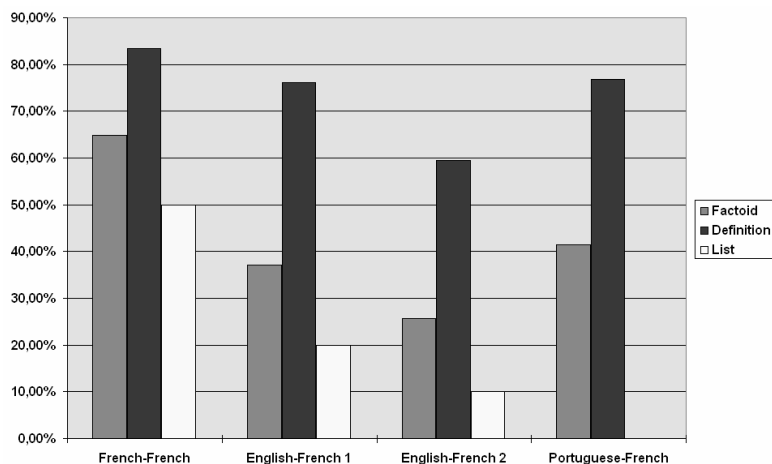


Fig. 3. Results of our 4 runs for each question type

As last year, our system is highly performing for questions of the type « Definition ». It is to be noted that this type of question, the “loss” of performance in a multi-lingual context, is less than for the other types of questions. This is due to the fact the “Definition” type of question relates most of the times, to acronyms or surnames of people, for which the translation is simpler and less ambiguous. The contribution of the Portuguese-French proper names lexicon seems obvious as the percentage of found definitions in this pair has moved from 68 % (CLEF 2005) to 77 % (CLEF 2006).

The « list-type » questions were only identified by our French module, thus none of this type of question were exact from the Portuguese, and the questions assessed as exact for English-French were, in fact, lists of one single element. For French, the proportion of exactly identified lists was honourable (50%), but this type of question remains difficult to process by our system.

Specific developments for the NIL questions had been implemented in CLEF 2006. In monolingual mode, the improvement is spectacular since the “precision rate” is now 0.56 versus 0.23 in 2005 and the “recall rate” 0.66 versus 0.25. For English-French they were respectively 0.29 v 0.14 and 0.66 v 0.30. Finally, for Portuguese-French, the “precision rate” was 0.26 v 0.13 and “recall rate” was 0.70 v 0.15.

Figure 4 presents statistics for answers evaluated as 'R' that stands for right. But CLEF proposed two other qualifications for answers that is 'U' for unjustified and 'X' for inexact. We think 'U' and 'X' answers would be often accepted by users, even 'X' answers if they are presented with their context. For question 57 *Qui est Flavio Briatore ? (Who is Flavio Briatore?)*, the answer provided by our system was *directeur général de Benetton Formula (general manager of Benetton Formula)*, whereas the awaited answer was *directeur général de Benetton Formula 1 (general manager of Benetton Formula 1)*. Likewise, for question 96 *A quel parti politique Jacques Santer appartient-il ? (Which political party does Jacques Santer belong to ?)*, the provided

answer by Qristal was *Parti chrétien-social dès 1966* (*Christian Social Party since 1966*) whereas the awaited answer was *Parti chrétien-social* (*Christian Social party*). This lead us to consider statistics for all answers considered as "not wrong", that is right (R), unjustified (U) or inaccurate (X):

Then we had a closer look to questions where the monolingual process finds the answer but the cross language does not. This leads us to the following remark. Questions are defined by reading the corpus and, deliberately or not, people formulating questions tend to reuse words or expressions mentioned in the text of the identified answer. On one hand, this influences the capacity of the system and the importance of each module in the overall process. For example, the use of synonyms is not that important for CLEF as it normally is. On the other hand, for cross language question answering, translations can be fuzzy and potentially quite far from the targeted word or expression especially when one uses English as an intermediate language. In this way, translated words are quite often different from the terms mentioned by both the question and the answer.

For question 1 "*Qu'est-ce qu'Atlantis ?*" ("*What is Atlantis ?*"), the question is translated from the English sentence "*What is Atlantis ?*" but in fact, the word "Atlantis" is normally translated by "Atlantide" in French and "Atlantis" is only kept when you want to speak of the spatial shuttle.

More generally, in comparing the monolingual and multilingual results, one observes that the longer the question is, the less proper names are present, and the less the results are satisfying. The quality loss is estimated to about 15 % for the questions of the "definition" type, and near 50 % for factual questions with temporal anchor. On the 200 questions used for the evaluation, 17 had no proper names and no dates. The following table provides the results of our runs for these questions :

	FR-FR	EN-FR 1	EN-FR 2	PT-FR
18	R	W	R	W
24	R	R	R	R
59	W	W	W	W
64	R	W	W	W
79	R	W	W	W
100	R	W	W	X
104	W	W	W	W
109	W	W	W	W
117	R	W	W	W
118	R	W	W	W
133	X	W	W	W
144	W	R	R	R
164	R	W	W	W
166	R	R	R	R
188	W	W	W	W
189	R	R	W	R
199	W	W	W	W
	58 %	24 %	24 %	24 %

The table shows that for these questions which quality of the translation is a crucial issue, the results are heavily deteriorated. Question 144, for which only the monolingual run returns an error, was a NIL question for which the French-French module returns, in despite, an answer while the other modules are returning a NIL.

Priberam, the company responsible for the Portuguese module in our engine, participated in CLEF 2006 for Portuguese and Spanish evaluation tracks. It is interesting to note that they obtained results very similar to our results for the Portuguese monolingual run [1] [2] and have similar degradations of results from monolingual to multilingual runs.

5 Outlines

Our CLEF 2006 results are noticeably better than those of our CLEF 2005 campaign, furthermore if one consider that “list” questions had been added this year. Notable too is our English-French run, using our Italian partner’s English module, returned less good results in 2006 versus 2005 ((32,5% v 39,5%), this confirming the overall greater difficulty attached to the questions this year.

The following modifications have generated the following improvements:

- Revised processing of the NIL questions. Although this is of little interest to the user, often requesting replies even inexact ones, the revision was implemented for CLEF and its evaluations. The end result has been to reach far better “precision” and “recall” rates for this type of questions.
- Slightly improved translations and primarily the use of the multilingual lexicon of translated proper names and acronyms. Thanks to these dictionaries, the deterioration between the French-French and French-Portuguese pairs has dropped significantly ie 43 % (2005) to 31 % (2006) detailed in the following : (a – CLEF 2005) from 64 % in French-French and 36.5% in Portuguese-French to (b- CLEF 2006) 68% in French-French and 47% in Portuguese-French).
- Betterment of the resources and the algorithm to detect the type of the question.

In respect of the deployed efforts, the development costs engaged and the resources upgrading, the global improvement of the results is not astonishing. It seems that the algorithms used by our modules find their limits around 70% of satisfying answers ! However, noticing that only 20 % of the answers are marked “wrong” in French-French, one may think that a revision of the delimitation of the extracted replies could, in the near future, allow the reach of success rates around 80 %.

Numerous other developments and improvements have been incorporated into our system, specially initiated in the framework of our M-CAST European project, but are not visible into the CLEF assessed results. For instance, the speed to return an answer has greatly improved, from an usual 3 seconds in 2005 to less than 1 second in 2006. This was obtained thanks to a preliminary fast analysis of the index returned sentences, which screens more than 80% of the sentences with no pivot-words of the question, hence a very weak probability to contain an answer.

After the CLEF 2005 evaluation, we had identified a few leads for improvement of our system. A few of them have been implemented this year, but a lot remains to be

done. We have started the elaboration of a “knowledge base”, from a Web-based data extraction, and currently targeting geographical data (country, province, town names), the whole to be extended, in the forthcoming months, to people names and events. We still not take into account the presentation of the document, and the answer extraction should be revised as still too imprecise.

As for the next coming years, our technology and system should evolve considerably as our firm is a partner of the Franco-German Quaero project, taking in charge the Question Answering issue for the French and English languages. Within this project, in partnership with the firm –Exalead– having developed the search engine “Eponyme”, we should market a general consumer and a professional version of our system, on closed corpora but more rewardingly on the billion Web pages. In this respect, our strategy should evolve towards a semantic tagging of the Web pages with indexation of the tagged items, in view to find the answers to a question within a timeframe of $1/10^{\text{th}}$ or $2/10^{\text{th}}$ of a second.

6 Conclusion

Despite the introduction of the questions of the « list » type, more difficult to process than the « definition » or « factual » ones, our system improved its results both in mono and multilingual mode. For factual questions, QRISTAL returns around 70 % of exact answers in monolingual mode and almost 50 % in multilingual mode.

A fine-tuned analysis of the results shows that the quality of the answer extraction still can be markedly improved as nearly 10 % of the answers have been assessed as “inexact” (“X”) or non-justified (“U”) in French-French mode, while the proportion of answers qualified as “false” was barely above 20 % in the same mode.

In the time elapsing between the 2 campaigns, our firm has invested almost 4 man/years in its system, but most developments were carried out on other areas that the quality of the said system. This was true for the answer delivering speed, the multitask operations, the Web accessibility, and the new English language module. We consider the real system improvement as having requested 1 man/year, an important investment for a comparatively reduced improvement.

The incorporation of our system into a Web-focused search engine through the Quaero project, is to necessitate in the coming years, a comprehensive revision of our methods in view to return the most precise answers in less than $2/10^{\text{th}}$ of a second. So, the core of the syntactic and semantic analysis processing will be batch processed and will produce a semantic tagging of the Web pages (or of the documents in closed corpora), along with the indexation of these tags (named entities, possible answers per type of question, key-words, etc.). Furthermore, the presence of a knowledge base fed and updated permanently via the Web, should permit the development of verification procedures of the answers, hence reducing the related errors. These procedures are of the most importance as, beside the delivered improvements, they avoid the delivery of nonsensical or absurd answers that correspondingly generate a user’s suspicion on the reliability of the system.

At a time when the Question Answering Systems find their first business use in firms or for the general audience, it is vital for these QA systems to avoid replicating the errors made while introducing the first grammar- checking or voice recognition

systems on the markets. As the average quality of these later systems in their initial versions has largely deceived their users, to the extent that any further satisfying development did not compensate for the deception and are still viewed as “unusable and of no interest”!

Acknowledgments

The authors thank all the engineers and linguists that took part in the development of QRISTAL. They also thank the Italian company Expert System and the Portuguese company Priberam for allowing them to use their modules for question analysis in English and Portuguese. They finally thank the European Commission which supported and still supports our development efforts through TRUST and M-CAST projects, and our coordinator Christian Gronoff from Semiosphere.

Last but not least, authors thank Carol Peters, Danilo Giampiccolo and Christelle Ayache for the remarkable organization of CLEF.

References

- [1] Amaral, C., Laurent, D., Martins, A., Mendes, A., Pinto, C.: Design & Implementation of a Semantic Search Engine for Portuguese. In: Proceedings of the Fourth Conference on Language Resources and Evaluation (2004)
- [2] Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C.: Priberam’s question answering system for Portuguese. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [3] Ayache, C., Grau, B., Vilnat, A.: Campagne d’évaluation EQueR-EVALDA : Évaluation en question-réponse, TALN 2005, 6-10 juin 2005, Dourdan, France, tome 2. – Ateliers & Tutoriels, pp. 63–72 (2005)
- [4] Clarke, C.L.A., Cormack, G.V., Lynam, T.R.: Exploiting Redundancy in Question Answering. In: Proceedings of 24th Annual International ACM SIGIR Conference (SIGIR 2001), pp. 358–365 (2001)
- [5] Grau B.: L’évaluation des systèmes de question-réponse, Évaluation des systèmes de traitement de l’information, TSTI, pp. 77–98, éd. Lavoisier (2004)
- [6] Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Williams, J., Bensley, J.: Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In: Proceedings of The Twelfth Text Retrieval Conference (TREC 2003) (2002)
- [7] Jijkun, V., Mishne, G., de Rijke, M., Schlobach, S., Ahn, D., Müller, K.: The University of Amsterdam at QALEF 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 15–17. Springer, Heidelberg (2005)
- [8] Laurent, D., Varone, M., Amaral, C., Fuglewicz, P.: Multilingual Semantic and Cognitive Search Engine for Text Retrieval Using Semantic Technologies. In: First International Workshop on Proofing Tools and Language Technologies, Patras, Grèce (2004)
- [9] Laurent, D., Séguéla, P.: QRISTAL, système de Questions-Réponses, TALN 2005, 6-10 juin 2005, Dourdan, France, tome 1. –Conférences principales, pp. 53–62 (2005)

- [10] Laurent, D., Séguéla, P., Nègre, S.: Cross-Lingual Question Answering using QRISTAL for CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [11] Laurent, D.: Industrial concerns of a Question-Answering system ? EACL 2006, Workshop KRAQ, April 3 2006, Trento, Italia (2006)
- [12] Laurent, D., Séguéla, P., Nègre, S.: QA better than IR ? EACL 2006, Workshop MLQA'06, April 4 2006, Trento, Italia (2006)
- [13] Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question Answering Track. In: Working Notes of the Workshop of CLEF 2004, 15-17 september 2004, Bath (2004)
- [14] Monz, C.: From Document Retrieval to Question Answering, ILLC Dissertation Series 2003-4, ILLC, Amsterdam (2003)
- [15] Voorhees, E.M.: Overview of the TREC 2003 Question Answering Track, NIST, PP. 54–68 (2003), http://trec.nist.gov/pubs/trec12/t12_proceedings.html

CLEF2006 Question Answering Experiments at Tokyo Institute of Technology

E.W.D. Whittaker, J.R. Novak, P. Chatain, P.R. Dixon, M.H. Heie,
and S. Furui

Dept. of Computer Science,
Tokyo Institute of Technology,
2-12-1, Ookayama, Meguro-ku,
Tokyo 152-8552 Japan

{edw,novakj,pierre,dixonp,heie,furui}@furui.cs.titech.ac.jp

Abstract. In this paper we present the experiments performed at Tokyo Institute of Technology for the CLEF2006 Multiple Language Question Answering (QA@CLEF) track. Our approach to QA centres on a non-linguistic, data-driven, statistical classification model that uses the redundancy of the web to find correct answers. For the cross-language aspect we employed publicly available web-based text translation tools to translate the question from the source into the corresponding target language, then used the corresponding mono-lingual QA system to find the answers. The hypothesised correct answers were then projected back on to the appropriate closed-domain corpus. Correct and supported answer performance on the mono-lingual tasks was around 14% for both Spanish and French. Performance on the cross-language tasks ranged from 5% for Spanish-English, to 12% for French-Spanish. Our method of projecting answers onto documents was shown not to work well: in the worst case on the French-English task we lost 84% of our otherwise correct answers. Ignoring the need for correct support information the exact answer accuracy increased to 29% and 21% correct on the Spanish and French mono-lingual tasks, respectively.

1 Introduction

In this paper we describe how we applied our recently developed statistical, data-driven approach to question answering (QA) to the task of multiple language QA in the CLEF2006 QA evaluation. The approach that we used, described in detail in [13,14,15], uses a noisy-channel formulation of the QA problem and the redundancy of the web to answer questions. Our approach permits the rapid development of factoid QA systems in new languages contingent only on the availability of sufficient training data, as described in [16].

Our approach is substantially different to conventional approaches to QA although it shares elements of other statistical, data-driven approaches to factoid QA found in the literature [1,2,3,7,10,11,12]. More recently, similar approaches to the answer typing employed in our system have appeared [9] although these

still use linguistic notions that our approach eschews in favour of data-driven classifications. While our approach results in a small number of parameters that must be optimised to minimise the effects of data sparsity and to make the search space tractable, we largely remove the need for numerous ad-hoc weights and heuristics that are an inevitable feature of many rule-based systems. The software for each language-specific QA system is identical; only the training data differs. All model parameters are determined when the data is loaded at system initialisation; these typically take only a couple of minutes to compute and do not change in between questions. New data or system settings can therefore be easily applied without the need for time-consuming model re-training.

Many contemporary approaches to QA require the skills of native speaking linguistic experts for the construction of rules and databases used in the QA systems. In contrast, our method allows us to include all kinds of dependencies in a consistent manner, is fully trainable and requires minimal human intervention once sufficient data is collected. This was particularly important in the case of our participation in CLEF this year since although our French QA system was developed by a native French-speaker, our Spanish system was built by a student with a conversational level of Spanish learnt in school.

Our QA systems perform best when there are numerous (redundant) sentences that contain the correct answer. We therefore use web documents extensively in our systems. To obtain good performance it helps if the correct answer co-occurs more frequently with words from the question than other candidate answers of the same type do. This information provides the only context for differentiating answers of the same type. CLEF requires support information from a fixed document collection for each answer; our answers, therefore, must subsequently be projected on to the appropriate collection. Inevitably this is a lossy operation as will be discussed in Section 5 and also means we never attempt to predict “unanswerable” questions by giving “NIL” as an answer.

The rest of this paper is organised as follows: we first present a summary in Section 2 of our mathematical framework for factoid QA. We describe the experimental setup specific to the CLEF2006 evaluation in Section 3 and present the results on each task that were obtained in the evaluation in Section 4. A discussion and conclusion are given in Sections 5 and 6.

2 QA as Classification with Non-linguistic Features

When responding to questions in a real-world scenario, humans consider a wide range of different factors including the identity of the questioner, immediate environmental and social context, and previously asked questions, among other variables. While such factors are clearly relevant, and humans are adept at integrating such information, modelling such features in an offline mode poses great difficulties. As a first pass, therefore, we limit ourselves to modelling the most straightforward dependence: the probability of an answer A depending on the question Q ,

$$P(A \mid Q) = P(A \mid W, X), \quad (1)$$

where A and Q are considered to be a string of l_A words $A = a_1, \dots, a_{l_A}$ and l_Q words $Q = q_1, \dots, q_{l_Q}$, respectively. Here $W = w_1, \dots, w_{l_W}$ represents a set of features describing the “question-type” part of Q such as *when*, *why*, *how*, etc. while $X = x_1, \dots, x_{l_X}$ represents a set of features that describe the “information-bearing” part of Q i.e. what the question is actually about and what it refers to. For example, in the questions, “*Who is the oldest person in the world?*” and “*How old is the oldest person in the world?*” the question-types *who* and *how old* differ, while the information-bearing component, *the oldest person in the world*, does not change.

Finding the best answer $\hat{A}$ involves a search over all available A for the one which maximises the probability of the above model

$$\hat{A} = \arg \max_A P(A \mid W, X). \quad (2)$$

Given the correct probability distribution this is guaranteed to give us the optimal answer in a maximum likelihood sense. We don’t know this distribution, and it is still difficult to model but, using Bayes’ rule and making various simplifying, modelling and conditional independence assumptions (as described in detail in [13,14,15]) Equation (2) can be rearranged to give

$$\arg \max_A \underbrace{P(A \mid X)}_{\text{retrieval model}} \cdot \underbrace{P(W \mid A)}_{\text{filter model}}. \quad (3)$$

The $P(A \mid X)$ model is essentially a statistical language model that models the probability of an answer sequence A given a set of information-bearing features X . We call this model the *retrieval model* and do not examine it further (see [13,14,15] for more details).

The $P(W \mid A)$ model matches a potential answer A with features in the question-type set W . For example, it relates place names with *where*-type questions. In general, there are many valid and equiprobable A for a given W so this component can only re-rank candidate answers obtained by the retrieval model. We call this component the *filter model* and outline its function in the following section.

2.1 Filter Model

The question-type feature set $W = w_1, \dots, w_{l_W}$ is constructed by extracting n -tuples ($n = 1, 2, \dots$) such as *Where*, *In what* and *When were* from the question Q . A set of $|V_W| = 2522$ single-word features is extracted based on frequency of occurrence in our collection of example questions.

Modelling the complex relationship between W and A directly is non-trivial. We therefore introduce an intermediate variable representing classes of example questions-and-answers (q-and-a) c_e for $e = 1 \dots |C_E|$ drawn from the set C_E . In order to construct these classes, given a set E of example q-and-a, we then define a mapping function $f : E \mapsto C_E$ such that for each example q-and-a

t_j , where $j = 1 \dots |E|$, $f(t_j) = e$ maps t_j into a particular class c_e . Thus each class c_e may be defined as the union of all component q-and-a features from each t_j satisfying $f(t_j) = e$. Finally, to facilitate modelling we say that W is conditionally independent of c_e given A so that,

$$P(W | A) = \sum_{e=1}^{|C_E|} P(W | c_e^W) \cdot P(c_e^A | A). \quad (4)$$

where c_e^W and c_e^A refer respectively to the subsets of question-type features and example answers for the class c_e .

Assuming conditional independence of the answer words in class c_e given A , and making the modelling assumption that the j th answer word a_e^j in the example class c_e is dependent only on the j th answer word in A we obtain:

$$P(W | A) = \sum_{e=1}^{|C_E|} P(W | c_e^W) \cdot \prod_{j=1}^{l_{Ae}} P(a_e^j | a_j). \quad (5)$$

Since our set of example q-and-a cannot be expected to cover all the possible answers to questions that may be asked we perform a similar operation to that above to give us the following:

$$P(W | A) = \sum_{e=1}^{|C_E|} P(W | c_e^W) \prod_{j=1}^{l_{Ae}} \sum_{a=1}^{|C_A|} P(a_e^j | c_a) P(c_a | a_j), \quad (6)$$

where c_a is a concrete class in the set of $|C_A|$ answer classes C_A .

The system using the above formulation of filter model given by Equation (6) is referred to as model ONE. Systems using the model given by Equation (4) are referred to as model TWO. Only systems based on model ONE were used in the official CLEF2006 evaluation systems described in this paper.

3 Experimental Setup for CLEF2006

The CLEF2006 tasks we took part in are as follows: F-F, S-S, F-S, E-S, E-F, F-E and S-E where F=French, S=Spanish and E=English. For each task we submitted only one run and up to ten ranked answers for each question. No classification of a question as factoid, definition or list question was performed prior to answering a question. Therefore all questions were treated as factoid questions since the QA systems we developed were trained using only factoid questions and the features extracted from factoid questions.

For the two mono-lingual tasks (French and Spanish) questions were passed as-is to the appropriate mono-lingual system after minimal query normalisation. For the cross-language tasks questions were first translated into the target language using web-based text translation tools: Altavista's Babelfish [5] for French-Spanish and Google Translate [6] for all other combinations. The same

minimal query processing is performed as for mono-lingual questions and the question is then passed to the appropriate mono-lingual QA system.

The corpora where answers were intended to be found were not actually used for answering questions. However, once a set of answers to a question had been obtained the final step was to project the answers on to the appropriate corpus. We relied on Lucene [4] to determine the document which had the highest ranking when the answer and question were used as a Boolean query. Snippets were likewise determined using Lucene's summary function. If an answer could be found somewhere in the document collection the snippets were further filtered to ensure that the snippet always included the answer string.

Due to time limitations we implemented the French and Spanish systems using only model ONE (Equation (6)). Although we have implementations for English using both models ONE and TWO, for consistency we used the English system that implemented model ONE only. In the TREC2005 QA evaluation model TWO outperformed model ONE by approximately 50% relative. Although there was insufficient time to implement systems using model TWO for the main CLEF evaluation a Spanish model TWO system was implemented in time for the real-time QA exercise that was held during the CLEF conference.

4 Results

All tasks comprised a set of 190 factoid and definition questions and 10 list questions. Our QA systems produced a maximum of 10 answers to each question, for all tasks. In general the top 3 answers for the factoid/definition questions were assessed and all list answers were assessed for exactness and support.

In Table 1 a breakdown is given of the results obtained on the two mono-lingual (French and Spanish) tasks for factoid/definition questions with answers in first place and for all answers to list questions.

Table 1. Breakdown of performance on the French and Spanish mono-lingual tasks by type of question and assessment of answer.

Task	Factoid questions				List questions			
	Right	ineXact	Unsupp.	CWS	Right	ineXact	Unsupp.	P@N
S-S	26 (13.7%)	1	29	0.035	3	0	0	0.03
F-F	27 (14.2%)	12	12	0.142	9	2	0	0.09

A similar breakdown for the five¹ cross-language tasks is given in Table 2 for factoid/definition questions with answers in first place and for all answers to list questions.

At the CLEF2006 workshop in Alicante a novel evaluation was performed to assess the speed and performance of systems in a Spanish language real-time QA task. Each participant was given the same 20 questions to answer and the

¹ Note that the Spanish-French cross-language task was not run in CLEF2006.

Table 2. Breakdown of performance on the English, French and Spanish cross-language combinations by type of question and assessment of answer

Task	Factoid questions				List questions			
	Right	ineXact	Unsupp.	CWS	Right	ineXact	Unsupp.	P@N
E-F	19 (10.0%)	6	8	0.017	4	1	1	0.06
E-S	11 (5.8%)	0	10	0.005	1	0	0	0.01
F-E	7 (3.7%)	10	37	0.003	1	3	15	0.01
S-E	10 (5.3%)	11	34	0.008	0	1	18	0.00
F-S	22 (11.6%)	0	15	0.037	2	0	0	0.02

Table 3. Mean reciprocal rank (MRR) of both exact and inexact answers and timings of systems that participated in the Spanish real-time QA task. (Participants anonymized since results are not official).

Group	MRR (Exact)	MRR (Inexact)	Time (s)
Group A	0.41	0.41	549
Group B	0.35	0.35	56
Group C	0.30	0.30	1966
TokyoTech	0.38	0.48	5141
Group E	0.24	0.28	76

emphasis was on speed and accuracy of results. The mean reciprocal rank (MRR) of both exact and inexact answers for the five participants are shown in Table 3.

5 Discussion

There were four main factors in our submissions to CLEF2006 that were expected to have a large impact on performance: (1) the mismatch in time period between the document collection and the web documents used for answering questions; (2) the use of factoid QA systems to also answer definition and list questions; (3) the effect of the machine translation tools for the cross-language tasks; and (4) the projection method of mapping answers back on to the document collection.

The inevitable mismatch in time periods between web search documents and the 1994-1995 EFE corpus led to inconsistencies for a limited number of questions: “¿Quién es el presidente de Letonia?” (“Who is the president of Latvia?”) and “¿Quién es el secretario general de la Interpol?” (“Who is the secretary general of Interpol?”) the answers to which have since changed.

For the sake of speed and efficiency we did not treat list and definition questions differently from factoids even though this method has not proven effective in earlier evaluations, such as TREC2005. However, since list questions were not identified in this year’s CLEF question sets we thought our efforts were better spent on improving the factoid aspect of the evaluation. For definition questions the independence assumptions in model ONE result in poor answer typing unless the answer can be defined in one or two words.

The substantially lower answer accuracies (between 3.7% and 12.0%) obtained on the cross-language tasks where Babelfish and Google Translate were used for question translation were generally expected due to the well documented low quality of such translation tools. For example, when the English question “*In which settlement was a mass slaughter of Muslims committed in 1995?*” was translated into French “settlement” became “règlement”. Consequently, answers given for this question related to the French legal system rather than a location. It was deemed unlikely that the highest result that was obtained, for French-Spanish, was due to using Babelfish rather than Google Translate and was instead due more to the relative similarity of the two languages (see Section 5.1). In any case, further improvements in machine-translation techniques will almost certainly result in considerable improvements in our cross-language QA performance and multi-language combination experiments.

We were far more surprised and disappointed with the loss incurred by our method of projecting answers which reduced our set of correct answers by 47% and 31% on the Spanish and French mono-lingual tasks, respectively. Ignoring the need for correct support information the performance would increase to 29% and 21% correct on the Spanish and French mono-lingual tasks, respectively. In the worst case, on the French-English task, we lost 84% of our otherwise correct answers i.e. we would have obtained an exact answer accuracy of 23% if the support requirement were ignored. Previous experience with projection onto the AQUAINT collection for English language answers on the TREC2005 QA task using the algorithm included in the open-source Aranea system [8] had shown fairly consistent losses of around only 20% although admittedly the algorithm that we applied in CLEF2006 was far simpler than that employed by Aranea.

In the next two sections we present brief language-specific discussions of the results that were obtained.

5.1 Spanish

As indicated in Table 2 results for the mono-lingual Spanish task were considerably better than those obtained for the cross-language tasks (English-Spanish, French-Spanish). The discrepancy between the results for the mono-lingual and cross-language Spanish test sets can be generally explained in terms of the relative accuracy of the automatic translation tools used as an intermediate step to obtaining results for the cross-language questions. Furthermore, the difference between the results for the French-Spanish task and those for the English-Spanish task can be attributed to the relative closeness of the language pair, with Spanish and French both being members of the Romance family of languages, rather than the use of different automatic translation tools. These differences aside, results for all three Spanish tasks exhibited similar characteristics.

The results on all three tasks that included Spanish as a source or target language were by far the best for factoid questions, especially those whose answers could be categorised as names or dates. Of the 26 exactly correct and supported answers obtained for the Spanish mono-lingual task, a total of 20 consisted of proper names or dates, (11 dates, 9 proper names). If the 29 correct but

unsupported answers are also taken into account, this total rises to 41, and accounts for approximately 75% of all correct answers obtained for this particular task.

Definition questions, in addition to being ambiguous in the evaluation sense, are much more difficult than factoids for our QA system to answer. Yet, definition question results were still better than we initially anticipated. These results included 2 exactly correct and supported answers, and 8 correct but unsupported answers in the definition category. A cursory analysis of the data revealed that each of these correct answers could be construed as the result of a categorisation process, whereby the subject of the question had been classified into a larger category, and this category was then returned as the answer, e.g. “¿Qué es la Quinoa?” (“What is Quinoa?”) answer: [a] cereal, and “¿Quién fue Alexander Graham Bell?” (“Who was Alexander Graham Bell?”) answer: [an] inventor. The system gives these ‘category’ words high scores due to the fact that they often appear in the context of proper nouns, where they are used as definitions, or as noun-qualifiers. However, because the system uses no explicit linguistic typology or categories, this results in occasional mismatches such as: “¿Quién es Nick Leeson?” (“Who is Nick Leeson?”) (answer: Barings). This answer would be categorised as a retrieval error since ‘Barings’ is a valid answer type for a Who-question but its high co-occurrence with the subject of the question results in an overly high retrieval model score.

5.2 French

For the French mono-lingual task, unsupported answers were not as much an issue as for the Spanish mono-lingual task, although there were still 12 unsupported answers for the factoid questions, 10 or 11 of which would have also been exact. For the English-French task, there were 8 unsupported answers for factoid questions, almost all of which were also exact. Projection onto French documents, however imperfect, seems to have been less of a problem than for the other languages though it is unlikely that the differences are significant.

Out of our 27 correct and supported answers on the mono-lingual task, 23 were places, dates or names. For those types of questions the answer types that were returned were usually correct. Questions involving numbers, however, were a serious problem: out of 15 “How many...” questions we got only one correct, and the answer types which were given were not consistent instead being dates, numbers, names or nouns. The same observation holds for “How old was X when...” questions, which were all incorrectly answered, with varying answer types for the answers given. With a rule-based or regular-expressions-based system it is difficult to make such errors. However, with our probabilistic approach, in which no hard decisions are made and all types of answers are valid but with varying probabilities, it is entirely possible to incur such filter model errors. Although some cases would be trivially remedied with a simple regular-expression this is against our philosophy; instead we feel the problem should be solved through better parameter estimation and better modelling of the training data, rather than ad-hoc heuristics.

Another interesting observation on the mono-lingual task was that for 19 questions where the first answer was inexact, wrong, or unsupported we got an exact and supported answer in second place. For answers in third place the number of exact and supported answers was only 3. In most of these cases, the answer types were the same. This is untypical of our results obtained previously on English and Japanese where there is typically a significant drop in the number of correct answers at each increase in the rank of answers considered.

The Spanish system, like the French system, was developed in a very short period of time. Making further refinements, increasing the amount of training data used, and implementing model TWO are expected to bring accuracy into line with the English system. The possible advantages of applying a refined cross-language approach to mono-lingual tasks, e.g. using the combined results of multiple mono-lingual systems to answer questions in a particular language, are also being investigated. This will provide a means of further exploiting the redundancy of the web, as well as a method to improve the results for languages which are still under-represented on the web.

On a positive note, although our system was the slowest in the Spanish QA real-time task, we achieved the second best MRR for exact answers and by far the highest MRR for inexact answers. Since system speed is more of an implementation issue (and dominated by retrieval and processing of web documents in our current system) we know speed can easily be improved. These results clearly show our system's potential for web-based factoid QA particularly when support information is not being assessed.

6 Conclusion

With the results obtained in the CLEF2006 QA evaluation we feel we have proven the language independence and generality of our statistical data-driven approach. Comparable performance using model ONE has been obtained under evaluation conditions on French and Spanish, and on English in both this evaluation and TREC2005/6. Post-evaluation experimentation with Japanese [15] has confirmed the efficacy of the approach for an Asian language as well.

State-of-the-art linguistics-based systems still achieve somewhat better absolute performance, however, steady improvement in various competitions such as TREC2006, and the comparatively brief two month development period for the Spanish and French systems with no need for a linguistic expert, continue to affirm the promise of our data-driven, non-linguistic approach to natural language QA.

Further work will concentrate on how to answer questions using less redundant data through data-driven query expansion methods and also look at removing the independence assumptions made in the formulation of the filter model to improve question and answer typing accuracy. We expect that improvements made on language-specific systems will feed through to improvements in all systems and

we hope to be able to compete in more and different language combinations in CLEF evaluations in the future.

7 Online Demonstration

A demonstration of the system using model ONE supporting questions in English, Japanese, Chinese, Russian, French, Spanish and Swedish can be found online at <http://www.inferret.com/>

Acknowledgements

This research was supported by the Japanese government's 21st century COE programme: "Framework for Systematization and Application of Large-scale Knowledge Resources".

References

1. Berger, A., Caruana, R., Cohn, D., Freitag, D., Mittal, V.: Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece (2000)
2. Brill, E., Dumais, S., Banko, M.: An Analysis of the AskMSR Question-answering System. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2002)
3. Echihiabi, A., Marcu, D.: A Noisy-Channel Approach to Question Answering. In: Proceedings of the 41st Annual Meeting of the ACL (2003)
4. Gospodnetic, O., Hatcher, E.: Lucene in Action. Manning (2005)
5. <http://babelfish.altavista.com>
6. <http://translate.google.com>
7. Ittycheriah, A., Roukos, S.: IBM's Statistical Question Answering System—TREC-11. In: Proceedings of the TREC 2002 Conference (2002)
8. Lin, J., Katz, B.: Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques. In: Proceedings of Twelfth International Conference on Information and Knowledge Management (CIKM 2003) (2003)
9. Pinchak, C., Lin, D.: A Probabilistic Answer Type Model. In: European Chapter of the ACL, Trento, Italy (2006)
10. Radev, D., Fan, W., Qi, H., Wu, H., Grewal, A.: Probabilistic Question Answering on the Web. In: Proc. of the 11th international conference on WWW, Hawaii, US (2002)
11. Ravichandran, D., Hovy, E., Josef Och, F.: Statistical QA—Classifier vs. Re-ranker: What's the difference? In: Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering (2003)
12. Soricut, R., Brill, E.: Automatic Question Answering: Beyond the Factoid. In: Proceedings of the HLT/NAACL 2004: Main Conference (2004)
13. Whittaker, E., Chatain, P., Furui, S., Klakow, D.: TREC2005 Question Answering Experiments at Tokyo Institute of Technology. In: Proceedings of the 14th Text Retrieval Conference (2005)

14. Whittaker, E., Furui, S., Klakow, D.: A Statistical Pattern Recognition Approach to Question Answering using Web Data. In: Proceedings of Cyberworlds (2005)
15. Whittaker, E., Hamonic, J., Furui, S.: A Unified Approach to Japanese and English Question Answering. In: Proceedings of NTCIR-5 (2005)
16. Whittaker, E., Hamonic, J., Yang, D., Klingberg, T., Furui, S.: Monolingual Web-based Factoid Question Answering in Chinese, Swedish, English and Japanese. In: Workshop on Multi-lingual Question Answering (EACL), Trento, Italy (2006)

Quartz: A Question Answering System for Dutch

David Ahn, Valentin Jijkoun, Joris van Rantwijk,
Maarten de Rijke, and Erik Tjong Kim Sang

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{ahn,jijkoun,rantwijk,mdr,erikt}@science.uva.nl
<http://ilps.science.uva.nl>

Abstract. We describe a question answering system for Dutch that we used for our participation in the 2006 CLEF Question Answering Dutch monolingual task. We give an overview of the system and focus on the question classification module, the multi-dimensional markup for data storage and access, the answer processing component, and our probabilistic approach to estimating answer correctness.

1 Introduction

For our earlier participations in the CLEF question answering track (2003–2005), we had developed a question answering architecture that uses different competing strategies to find answers in a document collection. For the 2005 edition of CLEF-QA, we focused on converting our text resources to XML in order to facilitate a QA-as-XML-retrieval strategy. This paper describes the 2006 edition of our QA system *Quartz*. We focused on converting all of our data resources (text and linguistic annotations) to fit in a database that uses XML-based multi-dimensional markup to provide uniform access to all annotated information. Additionally, we devoted attention to improving known weak parts of our system: question classification, answer clustering, and answer candidate score estimation.

This paper is divided in eight sections. In section 2, we give an overview of the current system architecture. In the following four sections, we zoom in on recent developments in the system: question classification (section 3), storing data with multi-dimensional markup (section 4), and probabilistic answer processing (sections 5 and 6). In section 7 we present the results of the CLEF QA 2006 evaluation of our system. We conclude in section 8.

2 System Description

The architecture of our Quartz QA system is an expanded version of a standard QA architecture consisting of parts dealing with question analysis, information retrieval, answer extraction, and answer post-processing (clustering, ranking, and selection). The Quartz architecture consists of multiple answer extraction modules, or *streams*, which share common question and answer processing components. These streams can be divided into three groups based on the corpus

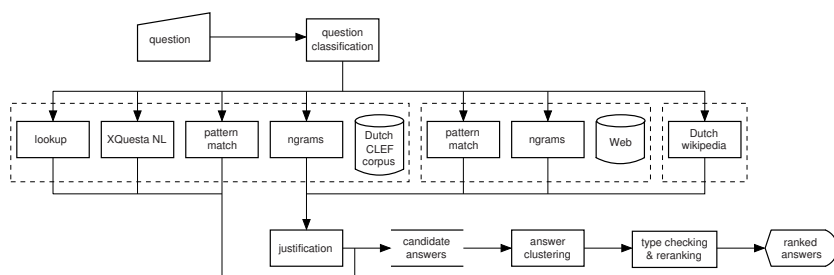


Fig. 1. Quartz: the University of Amsterdam’s Dutch Question Answering System

they employ: the CLEF-QA corpus, the Web, or Wikipedia. We describe these streams in the rest of this section. The common question and answer processing parts will be detailed in later sections on the architecture updates.

The Quartz system (Figure 1) contains five answer generating streams. The *Table Lookup* stream searches for answers in specialized knowledge bases which are extracted from the corpus offline (prior to question time) by manually defined rules. These rules take advantage of the fact that certain answer types, such as birthdays, are typically expressed in one of a small set of easily identifiable ways. Currently, the system makes use of 15 specific knowledge bases, containing information about locations, persons, monetary units, measurements, etc. Our question analysis module (section 3) selects which fields in which knowledge bases should be consulted for an incoming question and selects question keywords that are used by the *Table Lookup* stream to retrieve a list of candidate answers.

The *Ngrams* stream implements an approach similar to [1]. It looks for answers by extracting the most frequent word ngrams occurring in passages retrieved either from the document collection or from the Web (using Google). Similarly, the *Pattern Match* stream obtains passages relevant to the topic of the question either from the collection or from the Web and then searches for patterns formulated as regular expressions; the latter are generated by means of a hand-coded list of rules that use the result of the question analysis.

The *Wikipedia* stream relies on the structure of the Dutch Wikipedia. The focus of the question—usually the main named entity—is identified and looked up in Wikipedia. We use the standardized structure of Wikipedia to obtain a list of sentences relevant to the question focus. Then, the patterns of the *Pattern Match* stream and NE extraction are used to produce answer candidates. The *Justification* module (used also in *Ngrams* and *Web Pattern Match* streams) tries to find support for answer candidates in the Dutch CLEF corpus using IR.

The most advanced of the five CLEF-QA corpus streams is *XQesta*. It performs XPath queries against an XML version of the CLEF-QA corpus which contains the corpus text and additional annotations, including information about part-of-speech, syntactic chunks, named entities, temporal expressions, and dependency parses (from the Alpino parser [2]). *XQesta* retrieves passages of at least 400 characters, starting and ending at paragraph boundaries.

3 Question Classification

An important problem identified in the error analysis of our previous CLEF-QA results was a mismatch between the type of the answer and the type required by the question. To address this issue, we collected training data for the question classifier. We used 600 questions from the three previous CLEF QA tracks, 1000 questions from a Dutch trivia game and 279 questions from other sources—most importantly, the questions provided by users of our online demo. All the questions were manually classified by a single annotator.

Our question classification scheme assigns several classes to a question. The main reason is that while a coarse-grained type such as **person** would often be sufficient, for answering *which*-questions a fine-grained type such as **American President** would be needed. We used three different types of question classes: a *table type* that linked the question to a particular column in our Table Lookup knowledge bases (17 classes), a *coarse-grained type* that linked the question to the types recognized by our named-entity recognizer (7 classes), and a *fine-grained type* that linked the question to WordNet synsets (166 classes).

In Quartz, question classification is performed by a machine learner that represents a question using 10 features, such as the question word, the main verb of the question, the first noun following the question word, the top hypernym of the noun according to our type hierarchy, etc. In order to determine the top hypernym of the main noun of the question we used EuroWordNet [3]. With this small set of features the classifier achieved reasonable performance: 90% correct predictions for the coarse-grained classes.

Our classifier uses the memory-based learner Timbl [4]. We performed feature selection in order to identify the best subset of features for the classification task (using bidirectional hill-climbing, see [5]).

4 Multi-dimensional Markup

Our system makes use of several kinds of linguistic analysis tools, including a POS tagger, a named entity recognizer and a dependency parser [2]. These tools are run offline on the entire corpus, before any questions are posed.

The output of an analysis tool takes the form of a set of annotations: the tool identifies text regions in the corpus and associates some metadata with each region. Ideally, we would store these annotations as XML in order to access them through the powerful XQuery language. However, storing and accessing the annotations produced by different tools is not straightforward because the tools may produce conflicting regions. For example, a named entity may partially overlap with a phrasal constituent in such a way that the corresponding XML elements cannot be properly nested. It is thus not possible, in general, to construct a single XML tree that contains annotations from all tools that we used.

To deal with this problem, we have developed a general framework for the representation of multi-dimensional XML markup [6]. This framework stores

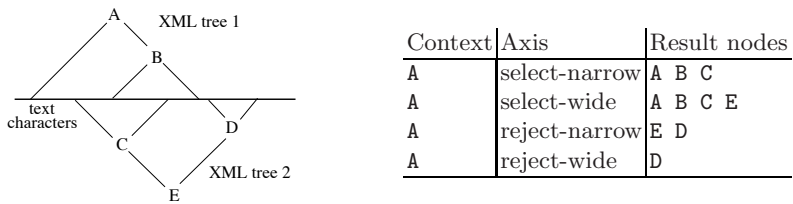


Fig. 2. Example document with two annotation layers

multiple layers of annotations referring to the same base document. Through an extension of the XQuery language, it is possible to retrieve information from several layers with a single query.

4.1 Stand-Off XML

We store all annotations as stand-off XML. Textual content is stripped from the XML tree and stored separately in a *blob* file (binary large object). Two *region attributes* are added to each XML element to specify the byte offsets of its start and end position with respect to the blob file. This process can be repeated for all annotation layers, producing identical blobs but different stand-off XML markup.

Figure 2 (left) shows an example of two annotation layers on the same text content. The stand-off markup in the two XML trees could be:

```
<A start="10" end="50">
  <B start="30" end="50"/>
</A>
<E start="20" end="60">
  <C start="20" end="40"/>
  <D start="55" end="60">
</E>
```

We use a separate system, called XIRAF [6], to coordinate the process of automatically annotating the corpus. XIRAF combines multiple text processing tools, each having an input descriptor that defines the format of the input (e.g., entire document or separate sentences), and a tool-specific wrapper that converts the tool output into stand-off XML annotation. We refer to [6] for a more detailed description of XIRAF.

4.2 Extending MonetDB/XQuery

The merged XML documents are indexed using MonetDB/XQuery [7], an XML database engine with full XQuery support. Its XQuery front-end consists of a compiler which transforms XQuery programs into a relational query language internal to MonetDB.

We extended the XQuery language by defining four new path axes that allow us to step between layers. The new axis steps relate elements by region overlap in the blob, rather than by nesting relations in the XML tree: **select-narrow** selects elements that have their region completely contained within the context

element's region; **select-wide** selects elements that have at least partial overlap with the context element's region; **reject-narrow** and **reject-wide** select the non-contained and non-overlapping region elements, respectively. The table in Figure 2 demonstrates the results of each of the new axes when applied to our example document.

In addition to the new axes, we added an XQuery function `so-blob($node)`. This function takes an element and returns the contents of that element's blob region. This function is necessary because all text content has been stripped from the XML markup, making it impossible to retrieve text directly from the XML tree.

The XQuery extensions were implemented by modifying the front-end of MonetDB/XQuery. An index on the offset attributes is used to make the new axis steps efficient even for large documents.

4.3 Using Multi-dimensional Markup for QA

Two streams in our QA system have been adapted to work with multi-dimensional markup: the *Table Lookup* stream and *XQuesta*. The table stream relies on a set of tables that are extracted from the corpus offline according to predefined rules. These extraction rules were rewritten as XQuery expressions and used to extract the tables from the collection text.

Previous versions of XQuesta had to generate separate XPath queries to retrieve annotations from several markup layers. Information from these layers had to be explicitly combined within the stream. Moving to multi-dimensional XQuery enables us to query several annotation layers jointly, handing off the task of combining the layers to MonetDB. Since XQuery is a superset of XPath, previously developed query patterns could still be used in addition to the new, multi-dimensional ones.

5 A Probabilistic Approach to Answer Selection

One major consequence of the multi-stream architecture of the Quartz QA system is the need for a module to choose among the candidate answers produced by the various streams. The principal challenge of such a module is making sense of the confidence scores attached by each stream to its candidate answers. This year, we made use of data from previous CLEF-QA campaigns in order to estimate correctness probabilities for candidate answers from stream confidence scores. Furthermore, we also estimated correctness probabilities conditioned on well-typedness in order to implement type-checking as Bayesian update. In the rest of this section, we describe how we estimated these probabilities. We describe how we use these probabilities in the following section.

5.1 From Scores to Probabilities

Each stream of our QA system attaches confidence scores to the candidate answers it produces. While these scores are intended to be comparable for answers

produced by a single stream, there is no requirement that they be comparable across streams. In order to make it possible for our answer re-ranking module (described in §6) to rank answers from different streams, we took advantage of answer patterns from previous editions of CLEF QA to estimate the probability that an answer from a given stream with a given confidence score is correct.

For each stream, we ran the stream over the questions from the previous editions of CLEF and binned the candidate answers by confidence score into 10 equally-sized bins. Then, for each bin, we used the available answer patterns to check the answers in the bin and based on these assessments, computed the maximum likelihood estimate for the probability that an answer with a score falling in the range of the bin would be correct. With these probability estimates, we can now associate with a new candidate answer a correctness probability based on its confidence score.

5.2 Type Checking as Bayesian Update

Type checking can be seen as a way to increase the information we have about the possible correctness of an answer. We discuss in section 6.2 how we actually type-check answers; in this section, we explain how we use Bayesian update to incorporate the results of type-checking into our probabilistic framework.

Given the prior probability of correctness for a candidate answer, $P(\textit{correct})$ (in our case, the MLE corresponding with the stream confidence score), as well as the information that it is well- or ill-typed (represented as the value of the random variable $\textit{well\_typed}$), we compute $P(\textit{correct} \mid \textit{well\_typed})$, the updated probability of correctness given well- (or ill-)typedness as follows:

$$P(\textit{correct} \mid \textit{well\_typed}) = P(\textit{correct}) \times \frac{P(\textit{well\_typed} \mid \textit{correct})}{P(\textit{well\_typed})} \quad (1)$$

In other words, the updated correctness probability of a candidate answer, given the information that it is well- or ill-typed, is the product of the prior probability and the ratio $P(\textit{well\_typed} \mid \textit{correct}) / P(\textit{well\_typed})$.

We estimate the required possibilities by running our type-checker on assessed answers from CLEF-QA 2003, 2004, and 2005. For question types in which type-checking is actually possible, the ratio for well-typed answers is 1.25. For ill-typed answers the ratio is 0.34.

6 Answer Processing

The multi-stream architecture of Quartz embodies a high-recall approach to question answering—the expectation is that using a variety of methods to find a large number of candidate answers should lead to a greater chance of finding correct answers. The challenge, though, is choosing *correct* answers from the many candidate answers returned by the various streams. The answer processing module described in this section is responsible for this task.

```

1: procedure ANSWERPROCESSING(candidates, question, tc)
2:   clusters  $\leftarrow$  CLUSTER(candidates)
3:   for cluster  $\in$  clusters do
4:     for ans  $\in$  cluster do
5:       ans.well_formed, ans.well_typed  $\leftarrow$  CHECK(ans, question)
6:     end for
7:      $P(\textit{cluster}) \leftarrow 1 - \prod_{\textit{ans} \in \textit{cluster}} (1 - P(\textit{ans}|\textit{ans.well\_typed}))$ 
8:     cluster.rep_answer  $\leftarrow \operatorname{argmax}_{\textit{ans} \in \{a \mid a \in \textit{cluster} \wedge a.\textit{well\_formed}\}} \text{length}(\textit{ans})$ 
9:   end for
10:  ranked_clusters  $\leftarrow \text{sort}_P(\textit{clusters})$ 
11: end procedure

```

Fig. 3. High-level answer processing algorithm

The algorithm used by the Quartz answer processing module is given in Figure 3. Candidate answers for a question are clustered (line 2: section 6.1), and each cluster is evaluated in turn (lines 3–9). Each answer in a cluster is checked for well-formedness and well-typedness (line 5: section 6.2). The correctness probability of the cluster $P(\textit{cluster})$ (line 7) is the combination of the updated correctness probabilities of the answers in the cluster (the prior correctness probabilities $P(\textit{ans}|\textit{ans.well\_typed})$ are described in section 5.2). The longer of the well-formed answers in the cluster is then chosen as the representative answer for the cluster (line 8). Finally, the clusters are sorted according to their correctness probabilities (line 10).

6.1 Answer Clustering

Answer candidates with similar or identical answer strings are merged into clusters. In previous versions of our system, this was done by repeatedly merging pairs of similar answers until no more similar pairs could be found. After each merge, the longest of the two answer strings was selected and used for further similarity computations.

This year, we moved to a graph-based clustering method. Formulating answer merging as a graph clustering problem has the advantage that it better captures the non-transitive nature of answer similarity. For example, it may be the case that the answers *oorlog* and *Wereldoorlog* should be considered similar, as well as *Wereldoorlog* and *wereldbeeld*, but not *oorlog* and *wereldbeeld*. To determine which of these answers should be clustered together, it may be necessary to take into account similarity relations with the rest of the answers.

Our clustering method operates on a matrix that contains a similarity score for each pair of answers. The similarity score is an inverse exponential function of the edit distance between the strings, normalized by the sum of the string lengths. The number of clusters is not set in advance but is determined by the algorithm. We used an existing implementation of a spectral clustering algorithm [8] to compute clusters within the similarity graph. The algorithm starts by putting all answers in a single cluster, then recursively splits clusters according to spectral

analysis of the similarity matrix. Splitting stops when any further split would produce a pair of clusters for which the normalized similarity degree exceeds a certain threshold. The granularity of the clusters can be controlled by changing the threshold value and the parameters of the similarity function.

6.2 Checking Individual Answers

The typing and well-formedness checks performed on answers depends primarily on the expected answer type of the question. Real type-checking can only be performed on questions whose expected answer type is a named-entity, a date, or a numeric expression. For other questions, only basic well-formedness checks are performed. Note that the probability update ratios described in section 5.2 are only computed on the basis of answers that are type-checkable. For answers to other questions, we use the same ratio for ill-formed answers as we do for ill-typed answers (0.34), but we do not compute any update for well-formed answers (i.e., we update with a ratio of 1.0).

To check typing and well-formedness of an answer we use a set of heuristic rules created by analysing the past performance of the system. E.g., for abbreviation or monetary unit questions the answer should consist of one word, answers to measurement question should contain numerical information, etc. For questions expecting names and dates as answers, we run an NE or date tagger, as appropriate, on the justification snippet to verify that the candidate answer is a name (for well-formedness) and is of the correct type (for well-typedness). Then, the results of answer checking are boolean values for well-formedness and well-typedness, as well as the possibly edited answer string.

7 Results and Analysis

For the CLEF 2006 QA task we submitted two Dutch monolingual runs. The run `uams06Tn1n1` used the full system with all streams and final answer selection. The run `uams06Nn1n1` used the full system but without type-checking.

The question classifier performed as expected for coarse question classes: 86% correct compared with a score of 85% on the training data. For most of the classes, precision and recall scores were higher than 80%; the exceptions are `miscellaneous` and `number`. For assigning table classes, the score was much lower than for the training data: 56% compared with 80%.

Table 1 lists the assessment counts for the two University of Amsterdam runs for the question answering track of CLEF-2006. The two runs had 14 different top answers of which four were assessed differently. In 2005 our two runs contained a large number of inexact answers (28 and 29). We are happy about those numbers being lower in 2006 (4 and 4) and the most frequent problem, extraneous information added to a correct answer, has almost disappeared. However, the number of correct answers dropped as well, from 88 in 2005 to 41. This is at least partly caused by the fact that the questions were more difficult this year.

The differences in the evaluation of the two runs are minimal and do not allow us to make any conclusions of the effectiveness of the typechecking in

Table 1. Assessment of the two runs and of one run per question type

Run	Total	Right	Unsupported	Inexact	Wrong	% Correct
uams06Tnlnl	200	40	2	4	154	20%
uams06Nnlnl	200	41	3	4	152	21%
Assessment for run uams06Nnlnl per question type						
factoid	116	28	1	2	85	24%
definition	39	9	0	1	29	23%
temporally restricted	32	3	1	1	27	8%
non-list	187	40	2	4	141	21%
list	13	0	6	0	31	0%

the Dutch system. This is in contrast with our experience with the English version of the system [9], where we noticed a substantial improvement. More experiments are needed to understand what causes the lack of improvement for Dutch: the quality of the Dutch type-checking itself or our probabilistic approach to combining different evidence for answer candidates (section 5).

Like last year, the questions could be divided in two broad categories: questions asking for lists and questions requiring singular answers. The second category can be divided in three subcategories: questions asking for factoids, questions asking for definitions and temporally restricted questions. We have examined the uams06Tnlnl-run answers to the questions of the different categories in more detail (Table 1). Our system failed to generate any correct answers for the list questions. For the non-list questions, 21% of the top answers were correct. Within this group, the temporally restricted questions¹ (8% correct) posed the biggest challenge to our system. In 2005, we saw similar differences between factoid, definition and temporally restricted questions. At that time the difference between the last category and the first two could be explained by the presence of a significant number of incorrect answers which would have been correct in another time period. This year, no such answers were produced by our system.

8 Conclusion

We have described the fourth iteration of our system for the CLEF Question Answering Dutch mono-lingual track (2006). This year, our work has focused on converting all data repositories of the system (text, annotation and tables) to XML and allowing them to be accessed via a single interface. Additionally, we modified parts of our system which we had suspected of weaknesses: question classification and answer processing.

Our experiments with the English version of the system [10] show that if a fine-tuned answer selection module is employed, all QA streams contribute to performance: removing any single stream results in fewer answered questions. We still need to perform similar experiments for the Dutch system.

¹ The official assessments indicate only one temporally restricted question, which seems unlikely. We count all questions with time expressions as temporally restricted.

Acknowledgments

This research was supported by various grants from the Netherlands Organisation for Scientific Research (NWO). Valentin Jijkoun was supported under project numbers 220.80.001, 600.065.120 and 612.000.106. Joris van Rantwijk and David Ahn were supported under project number 612.066.302. Erik Tjong Kim Sang was supported under project number 264.70.050. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220.80.001, 264.70.050, 354.20.005, 600.065.120, 612.13.001, 612.000.106, 612.066.302, 612.069.006, 640.-001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

1. Dumais, S., Banko, M., Brill, E., Lin, J., Ng, A.: Web question answering: Is more always better? In: Bennett, P., Dumais, S., Horvitz, E. (eds.) *Proceedings of SIGIR'02*, pp. 291–298 (2002)
2. van Noord, G.: At last parsing is now operational. In: *Proceedings of TALN 2006*, Leuven, Belgium (2006)
3. Vossen, P.: *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht (1998)
4. Daelemans, W., Zavrel, J., van der Sloot, K., van den Bosch, A.: *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide*. University of Tilburg, ILK Technical Report ILK-0402 (2004), <http://ilk.uvt.nl/>
5. Caruana, R., Freitag, D.: Greedy attribute selection. In: *Proceedings of the Eleventh International Conference on Machine Learning*, New Brunswick, NJ, USA, pp. 28–36. Morgan Kaufman, San Francisco (1994)
6. Alink, W., Jijkoun, V., Ahn, D., de Rijke, M., Bonz, P., de Vries, A.: Representing and querying multi-dimensional markup for question answering. In: *Proceedings of the 5th Workshop on NLP and XML*, ACL, pp. 3–9 (2006)
7. CWI: MonetDB Website: <http://www.monetdb.nl/>
8. Dragone, L.: Spectral clusterer for WEKA (2002), <http://www.luigidragone.com/datamining/spectral-clustering.html>
9. Schlobach, S., Ahn, D., de Rijke, M., Jijkoun, V.: Data-driven type checking in open domain question answering. *Journal of Applied Logic* (2006)
10. Jijkoun, V., de Rijke, M.: Answer selection in a multi-stream open domain question answering system. In: McDonald, S., Tait, J. (eds.) *ECIR 2004*. LNCS, vol. 2997, Springer, Heidelberg (2004)
11. Jijkoun, V., de Rijke, M.: Retrieving answers from frequently asked questions pages on the web. In: *Proceedings of the Fourteenth ACM conference on Information and knowledge management (CIKM 2005)*, ACM Press, New York (2005)

Experiments on Applying a Text Summarization System for Question Answering

Pedro Paulo Balage Filho, Vinícius Rodrigues de Uzêda,
Thiago Alexandre Salgueiro Pardo, and Maria das Graças Volpe Nunes

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo
CP 668, 13.560-970 São Carlos-SP, Brasil
<http://www.nilc.icmc.usp.br>
{pedrobalage,vruzeda}@gmail.com, {tasparado,gracan}@icmc.usp.br

Abstract. In this paper, we present and analyze the results of the application of a text summarization system – GistSumm – to the task of monolingual question answering at CLEF 2006 for Portuguese texts. We hypothesized that topic-oriented summarization techniques could be able to produce more accurate answers. However, our results showed that there is a big gap to be overcome for summarization to be directly applied to question-answering tasks.

Keywords: Question answering, text summarization.

1 Introduction

We present and analyze in this paper the results of the application of a summarization system to the task of monolingual Question Answering (QA) at CLEF 2006 for Portuguese texts. We aimed at assessing the performance of the system in answering questions using its topic-oriented summarization method: each question was considered the topic around which the summary should be built, hopefully containing the appropriate answer. The idea is that summarization techniques may be able to produce more accurate answers, since the irrelevant information in texts is ignored.

The system we used is GistSumm [3], a summarizer with very high precision in identifying the main idea of texts, as indicated by its participation in DUC (Document Understanding Conference) 2003 evaluation.

Two runs were submitted to CLEF. For one run, we submitted the summarization system results without any post-processing step. For the other one, we applied a simple filter that we developed for finding in the produced summary a more factual answer. The performance of both methods at CLEF was very poor, indicating that simple summarization techniques alone are not enough for question answering tasks.

The summarization system we used and the filter we developed are briefly described in the next section. Our results at CLEF are reported in Section 3.

2 The System

GistSumm is an automatic summarizer based on an extractive method, called gist-based method. In order to make GistSumm to work, the following premises must hold: (a) every

text is built around a main idea, namely, its gist; (b) it is possible to identify in a text just one sentence that best expresses its main idea, namely, the gist sentence. Based on them, the following hypotheses underlie GistSumm methodology: (I) through simple statistics, the gist sentence or an approximation of it can be determined; (II) by means of the gist sentence, it is possible to build coherent extracts conveying the gist sentence itself and extra sentences from the source text that complement it.

The original GistSumm comprises three main processes: text segmentation, sentence ranking, and extract production. Text segmentation identifies the sentences in the text. Sentence ranking is based on the keywords method [2]: it scores each sentence of the source text by summing up the frequency of its words in the whole text. For producing summaries (i.e., generic summaries), the gist sentence is chosen as the one with the highest score. Extract production focuses on selecting other sentences from the source text to be included in the extract, based on: (a) gist correlation and (b) relevance to the overall content of the source text. A new version of GistSumm [1] is able to treat and summarize structured texts, e.g., scientific texts, producing an overall extract composed of smaller extracts of each text section. A characteristic of both systems is the possibility to produce summaries based on a specific topic suggested by the user. For this, the gist sentence is chosen as the one with the highest score in relation to the specified topic, with this correlation being measured by the cosine measure [4].

For participating at CLEF, we submitted two runs. For the first run, we simply returned, for each question, the highest scored gist sentence found in the topic-oriented summarization mode, with the question being the topic. In this manner we were searching for possible answers on the text that had better correlation with the question. For running CLEF experiments, the gist sentence was chosen independently for each section of each text in CLEF database. The best gist sentence was then selected to be the answer.

For the other run, we developed a filter for finding, for each question, a more restricted answer inside sentences previously indicated by GistSumm. These sentences were empirically set to be the 6 best scored ones in the topic-oriented summarization mode for the complete data collection, i.e., the 6 sentences in the whole text collection with the best correlations to the question. We forced the system to select sentences from different texts.

For each question, the filter works as follows:

- initially, it annotates the 6 selected sentences using a POS tagger;
- then, it tries to determine the type of the question posted by CLEF by analyzing its first words: if the question starts with “who”, then the filter knows that a person must be found; if it starts with “where”, then a place must be found; if it starts with “when”, then a time period must be found; if it starts with “how many”, then a quantity must be found; for any other case, the filter aborts its operation and returns as answer the same answer that would be given in the first run, i.e., the highest scored gist sentence;
- if the question type could be determined, then the filter performs a pattern matching process: if the question requires a date as answer, the filter will look for text spans in the 6 sentences that conforms, for example, to the pattern “month/day/year”; if a person or a place is required, then the filter will search for proper nouns (indicated by the POS tagger); if a quantity is required, the filter will search for expressions formed by numbers followed by nouns;

- if it was possible to find at least one answer in the last step, the first answer found is returned; otherwise, the answer is set to NIL, which ideally would indicate that the text collection do not contain the answer to the question.

The obtained results for both methods are reported in the next section.

3 Results and Discussion

We run our experiments on the Portuguese data collection for the following reasons: Portuguese is our native language and, thus, this enables us to better judge the results; Portuguese is one of the languages supported by GistSumm. CLEF Portuguese data collection contains news texts from Brazilian newspaper *Folha de São Paulo* and Portuguese newspaper *Público*, from years 1994 and 1995.

The QA track organizers made available 200 questions, which included the so called “factoid”, “definition”, and “list” questions. As defined by CLEF, factoid questions are fact-based questions, asking, for instance, for the name of a person or a location; definition questions are questions like “What/Who is X?”; list questions are questions that require a list of items as answers, for example, “Which European cities have hosted the Olympic Games?”. There were different amounts of each question type in QA track: 139 factoid questions, 47 definition questions, and 12 list questions.

The main evaluation measure used by CLEF is accuracy. For this measure, human judges had to tell, for each question, if the answer was right, wrong, unsupported (i.e., the answer contains a correct information but the provided text do not support it, or the text do not originate from the data collection), inexact (the answer contains a correct information and the provided text support it, but the answer is incomplete/truncated or is longer than the minimum amount of information required), or unassessed (for the case that no judgment was provided for the answer). Some variations of the accuracy measure are the Confidence Weighted Score (CWS), the Mean Reciprocal Rank Score (MRRS) and the K1 measure. For the interested reader, we suggest to refer to CLEF evaluation guidelines for these measures definitions. Table 1 and 2 show the results for the two submitted runs.

One can see that the results for both runs are very poor. The results were slightly better for the second run, given the use of the filter, even though the 3 right answers for the non-list questions originated from NIL answers. The performance of our system can be better visualized in the example (from CLEF data) in Figure 1. In this example, the system returned the 6 best correlated sentences with the question. Although the cosine measure is efficient when we desire to identify sentences of a topic to compose a summary, we can

Table 1. Results for the first run (without filtering the results)

Accuracy			CWS	MRRS	K1
Questions	Factoid and definition	List	0	0	-0.6445
Judgments					
Right	0	0			
Wrong	179	9			
Unsupported	7	0			
Inexact	2	3			
Unassessed	0	0			

Table 2. Results for the second run (filtering the results)

Accuracy			CWS	MRRS	K1
Questions	Factoid and definition	List	0.0002	0.0160	-0.5134
Judgments					
Right	3	0			
Wrong	179	10			
Unsupported	4	0			
Inexact	2	2			
Unassessed	0	0			

observe that it does not show to be good for question answering. In this example, we can also see that the filter was not applied because the question identifier (“What is the name”) wasn’t in its rules.

Analyzing the performance of the system for the entire set of questions and answers submitted, we could identify some important points to be investigated in the future:

- the sentence level analysis usually carried out by summarization systems are not sufficiently fine-grained to appropriately answer questions;
- although the cosine measure may be good for building topic-oriented summaries, it is not good for searching for answers, since it tends to select shorter sentences that have more common words with the question, ignoring longer sentences that may probably contain the answer;
- although the answer could not be evidenced by the best scored sentence, an analysis of the other sentences showed that the answer could be found among the 100 better scored sentences, in general;
- the filter we developed was too naive in trying to classify the questions in only 4 types (“who”, “where”, “when” and “how many”); many more variations were used in CLEF.

After CLEF experiments, we believe that simple summarization techniques are not enough for question answering tasks, even if these systems are good in what they do. We believe that much more work in this direction is needed.

Question: *Como se chama a primeira mulher a escalar o Evereste sem máscara de oxigênio?* (in English, What is the name of the first woman to climb Everest without oxygen mask?)

First selected answers and their cosine measure (in relation to the question):

Cosine	Answer
0.500000	<i>Ou: Uma mulher como eu</i> (Or: A woman like me)
0.500000	<i>Como se chama?</i> (What is her name?)
0.500000	<i>Como se chama?</i> (What is her name?)
0.472456	<i>Em Maio deste ano, Hargreaves, de 33 anos, tinha se tornado a primeira mulher a escalar o Evereste, sozinha e sem a ajuda de oxigênio.</i> (In May of this year, Hargreaves, 33 years, became the first woman to climb Everest, alone and without the oxygen aid.)

Selected answer for 1st run: *Ou: Uma mulher como eu* (Or: A woman like me)

Selected answer for 2nd run: *Ou: Uma mulher como eu* (Or: A woman like me)

Fig. 1. Example of an answer submitted to CLEF

Acknowledgments

The authors are grateful to CAPES, CNPq and FAPESP for supporting this work.

References

- [1] Balage Filho, P.P., Uzêda, V.R., Pardo, T.A.S., Nunes, M.G.V.: Estrutura Textual e Multiplicidade de Tópicos na Sumarização Automática: o Caso do Sistema GistSumm. Technical Report 283. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (2006)
- [2] Luhn, H.: The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2, 159–165 (1958)
- [3] Pardo, T.A.S., Rino, L.H.M., Nunes, M.G.V.: GistSumm: A Summarization Tool Based on a New Extractive Method. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.G.V. (eds.) *PROPOR 2003*. LNCS, vol. 2721, pp. 210–218. Springer, Heidelberg (2003)
- [4] Salton, G.: *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, Reading (1989)

N-Gram vs. Keyword-Based Passage Retrieval for Question Answering

Davide Buscaldi, José Manuel Gomez, Paolo Rosso, and Emilio Sanchis

Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi,jogomez,proso,esanchis}@dsic.upv.es

Abstract. In this paper we describe the participation of the Universidad Politécnica of Valencia to the 2006 edition, which was focused on the comparison between a Passage Retrieval engine (JIRS) specifically aimed to the Question Answering task and a standard, general use search engine such as Lucene. JIRS is based on *n*-grams, Lucene on keywords. We participated in three monolingual tasks: Spanish, Italian and French. The obtained results show that JIRS is able to return high quality passages, especially in Spanish.

1 Introduction

Most of the Question Answering (QA) systems that are currently used in the CLEF and TREC¹ QA exercises are based on common keyword-based Passage Retrieval (PR) methods [3,1,8,5]. QUASAR is a mono/cross-lingual QA system built around the JIRS PR system [6], that is based on *n*-grams instead of keywords. JIRS is specifically oriented to the QA task; it can also be considered as a language-independent PR system, because it does not use any knowledge about the lexicon and the syntax of the language during question and passage processing phases. The crucial step performed by JIRS is the comparison between the *n*-grams of the question and those of the passages retrieved by means of typical keyword-based search. The main objective of our participation to QA@CLEF2006 was the comparison of JIRS with a classical PR engine (in this case, Lucene²). In order to do this, we actually implemented two versions of QUASAR, which differ only for the PR system used. With regard to the improvements over last year's system, our efforts were focused on the Question Analysis module, which in contrast to the one used in 2005 does not use Support Vector Machines in order to classify the questions. Moreover, all modules are now better integrated in the complete system.

The 2006 CLEF QA task introduced some challenges with respect to the previous edition: list questions, the lack of a label distinguishing “definition” questions from other ones, and the introduction of another kind of “definitions”, that we named *object definitions*. This forced us to change slightly the class ontology we used in 2005.

¹ <http://trec.nist.gov>

² <http://lucene.apache.org>

In the next section, we describe the structure and the building blocks of our QA system. In section 3 we discuss the results of QUASAR in the 2006 CLEF QA task.

2 Architecture of QUASAR

The architecture of QUASAR is shown in Fig.1.

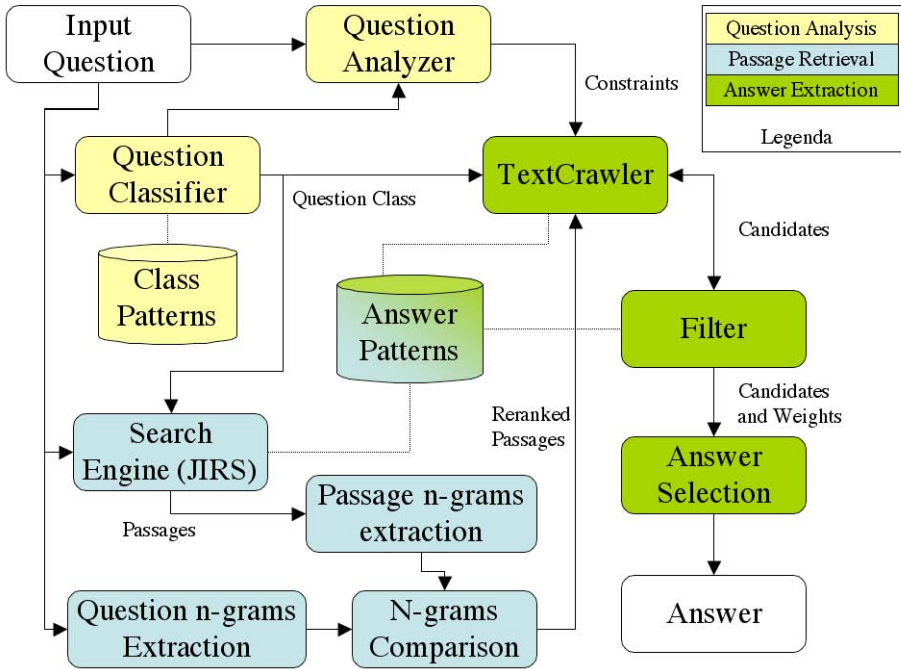


Fig. 1. Diagram of the QA system

Given a user question, this will be handed over to the *Question Analysis* module, which is composed by a *Question Analyzer* that extracts some constraints to be used in the answer extraction phase, and by a *Question Classifier* that determines the class of the input question. At the same time, the question is passed to the *Passage Retrieval* module, which generates the passages used by the *Answer Extraction* (AE) module together with the information collected in the question analysis phase in order to extract the final answer.

2.1 Question Analysis Module

This module obtains both the expected answer type (or *class*) and some constraints from the question. Question classification is a crucial step of the processing since the Answer Extraction module uses a different strategy depending on

Table 1. QC pattern classification categories

L0	L1	L2
NAME	ACRONYM PERSON TITLE FIRSTNAME LOCATION	COUNTRY CITY GEOGRAPHICAL
DEFINITION	PERSON ORGANIZATION OBJECT	
DATE	DAY MONTH YEAR WEEKDAY	
QUANTITY	MONEY DIMENSION AGE	

the expected answer type; as reported by Moldovan et al. [4], errors in this phase account for the 36.4% of the total number of errors in Question Answering.

The different answer types that can be treated by our system are shown in Table 1. We introduced the “FIRSTNAME” subcategory for “NAME” type questions, because we defined a pattern for this kind of questions in the AE module. We also specialized the “DEFINITION” class into three subcategories: “PERSON”, “ORGANIZATION” and “OBJECT”, which was introduced this year (e.g.: *What is a router?*). With respect to CLEF 2005, the Question Classifier does not use a SVM classifier.

Each category is defined by one or more patterns written as regular expressions. The questions that do not match any defined pattern are labeled with *OTHER*. If a question matches more than one pattern, it is assigned the label of the longest matching pattern (i.e., we consider longest patterns to be less generic than shorter ones).

The Question Analyzer has the purpose of identifying the constraints to be used in the AE phase. These constraints are made by sequences of words extracted from the POS-tagged query by means of POS patterns and rules. For instance, any sequence of nouns (such as *ozone hole*) is considered as a relevant pattern. The POS-taggers used were the SVMtool³ for English and Spanish, and the TreeTagger⁴ for Italian and French.

There are two classes of constraints: a *target* constraint, which is the word of the question that should appear closest to the answer string in a passage,

³ <http://www.lsi.upc.edu/nlp/SVMTool/>

⁴ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

and zero or more *contextual* constraints, keeping the information that has to be included in the retrieved passage in order to have a chance of success in extracting the correct answer. For example, in the following question: “*Dónde se celebraron los Juegos Olímpicos de Invierno de 1994?*” (*Where did the Winter Olympic games of 1994 take place?*) *celebraron* is the target constraint, while *Juegos Olímpicos de Invierno* and *1994* are the contextual constraints. There is always only one target constraint for each question, but the number of contextual constraint is not fixed. For instance, in “*Quién es Neil Armstrong?*” the target constraint is *Neil Armstrong* but there are no contextual constraints.

2.2 The JIRS Passage Retrieval Module

The passages containing the relevant terms are retrieved by JIRS using a classical keyword-based IR system. This year, the module was modified in order to rank better the passages which contain an answer pattern matching the question type. Therefore, this module is not as language-independent as in 2005 because it uses informations from the Question Classifier and the patterns used in the Answer Extraction phase.

Sets of unigrams, bigrams, ..., n -grams are extracted from the passages and from the user question. In both cases, n is the number of question terms. These n -gram sets are compared in order to obtain the weight of each passage, which is proportional to the size of the question n -grams found in the passage.

For instance, if the question is “*What is the capital of Croatia?*” and the system retrieves the following two passages: “*...Tudjman, the president of Croatia, met Eltsin during his visit to Moscow, the capital of Russia...*”, and “*...they discussed the situation in Zagreb, the capital of Croatia...*”. The second passage must have more importance because it contains the 4-gram “*the capital of Croatia*”, whereas the first one contains the 3-gram “*the capital of*” and the 1-gram “*Croatia*”. This example also shows the advantage of considering n -grams instead of keywords: the two passages contains the same question keywords, but only one of them contains the right answer.

In order to calculate the weight of n -grams of every passage, the greatest n -gram in the passage is identified and it is assigned a weight equal to the sum of all its term weights. Subsequently, smaller n -grams are searched. The weight of every term is determined by means of formula (1):

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \quad (1)$$

Where n_k is the number of passages in which the term appears, and N is the number of passages. We make the assumption that stopwords occur in every passage (i.e., $n_k = N$ for stopwords). Therefore, if the term appears once in the passage collection, its weight will be equal to 1 (the greatest weight).

Sometimes a term unrelated to the question can obtain a greater weight than those assigned to the Named Entities (NEs), such as names of persons, organizations and places, dates. The NEs are the most important terms of the question

and it does not make sense to return passages which do not contain them. Therefore, NEs are given a greater weight than the other question terms, in order to force its presence in the first ranked passages. NEs are recognized by simple rules, such as capitalization, or by checking if they are a number. Once all the terms have been weighted, the sum is normalized.

JIRS can be obtained at the following URL: <http://jirs.dsic.upv.es>.

2.3 Answer Extraction

The input of this module is constituted by the n passages returned by the PR module and the constraints (including the expected type of the answer) obtained through the *Question Analysis* module. A *TextCrawler* is instantiated for each of the n passages with a set of patterns for the expected type of the answer and a pre-processed version of the passage text. For CLEF 2006, we corrected some errors in the patterns and we also introduced new ones.

Some patterns can be used for all languages; for instance, when looking for proper names, the pattern is the same for all languages. The pre-processing of passage text consists in separating all the punctuation characters from the words and in stripping off the annotations of the passage. It is important to keep the punctuation symbols because we observed that they usually offer important clues for the individuation of the answer (this is true especially for *definition* questions): for instance, it is more frequent to observe a passage containing “*The president of Italy, Giorgio Napolitano*” than one containing “*The president of Italy IS Giorgio Napolitano*” ; moreover, movie and book titles are often put between apices.

The positions of the passages in which occur the constraints are marked before passing them to the TextCrawlers. A difference with 2005 is that now we do not use the Levenshtein-based spell-checker to compare strings in this phase now.

The TextCrawler begins its work by searching all the passage’s substrings matching the expected answer pattern. Then a weight is assigned to each found substring s , depending on the positions of s with respect to the constraints, if s does not include any of the constraint words. If in the passage are present both the target constraint and one or more of the contextual constraints, then the product of the weights obtained for every constraint is used; otherwise, it is used only the weight obtained for the constraints found in the passage.

The *Filter* module takes advantage of a mini knowledge base in order to discard the candidate answers which do not match with an allowed pattern or that do match with a forbidden pattern. For instance, a list of country names in the four languages has been included in the knowledge base in order to filter country names when looking for countries. When the Filter module rejects a candidate, the TextCrawler provide it with the next best-weighted candidate, if there is one.

Finally, when all TextCrawlers end their analysis of the text, the *Answer Selection* module selects the answer to be returned by the system. The following strategies apply:

- Simple voting (SV): The returned answer corresponds to the candidate that occurs most frequently as passage candidate.
- Weighted voting (WV): Each vote is multiplied for the weight assigned to the candidate by the TextCrawler and for the passage weight as returned by the PR module.
- Maximum weight (MW): The candidate with the highest weight and occurring in the best ranked passage is returned.
- Double voting (DV): As simple voting, but taking into account the second best candidates of each passage.
- Top (TOP): The candidate elected by the best weighted passage is returned.

We used the Confidence Weighted Score (CWS) to select the answer to be returned to the system, relying on the fact that in 2005 our system was the one returning the best values for CWS [7]. For each candidate answer we calculated the CWS by dividing the number of strategies giving the same answer by the total number of strategies (5), multiplied for other measures depending on the number of returned passages (n_p/N , where N is the maximum number of passages that can be returned by the PR module and n_p is the number of passages actually returned) and the averaged passage weight. The final answer returned by the system is the one with the best CWS. Our system always return only one answer (or NIL), although 2006 rules allowed to return more answers per question. The weighting of NIL answers is slightly different, since it is obtained as $1 - n_p/N$ if $n_p > 0$, 0 elsewhere.

The snippet for answer justification is obtained from the portion of text surrounding the first occurrence of the answer string. The snippet size is always 300 characters (150 before and 150 after the answer) + the number of characters of the answer string.

3 Experiments and Results

We submitted two runs for each of the following monolingual task: Spanish, Italian and French. The first runs (labelled *upv_061*) use the system with JIRS as PR engine, whereas for the other runs we used Lucene, adapted to the QA task with the implementation of a weighting scheme that privileges long passages and is similar to the word-overlap scheme of the MITRE system [2]. In Table 2 we show the overall accuracy obtained in all the runs.

With respect to 2005, the overall accuracy increased by $\sim 3\%$ in Spanish and Italian, and by $\sim 7\%$ in French. We suppose that the improvement in French is due to the fact that the target collection was larger this year. Spanish is still the language in which we obtain the best results, even if we are not sure about the reason: a possibility is that this can be due to the better quality of the POS-tagger used in the analysis phase for the Spanish language.

We obtained an improvement over the 2005 system in factoid questions, but also worse results in definition ones, probably because of the introduction of the *object* definitions.

Table 2. Accuracy results for the submitted runs. Overall: overall accuracy, factoid: accuracy over factoid questions; definition: accuracy over definition questions; nil: precision over nil questions (correctly answered nil/times returned nil); CWS: confidence-weighted score.

task	run	overall	factoid	definition	nil	CWS
es-es	upv_061	36.84%	34.25%	47.62%	0.33	0.225
	upv_062	30.00%	27.40%	40.48%	0.32	0.148
it-it	upv_061	28.19%	28.47%	26.83%	0.23	0.123
	upv_062	28.19%	27.78%	29.27%	0.23	0.132
fr-fr	upv_061	31.58%	31.08%	33.33%	0.36	0.163
	upv_062	24.74%	26.35%	19.05%	0.18	0.108

Table 3. Inter-agreement between the two systems, calculated by means of the AoA and AoRA measures

task	AoA	AoRA	Collection size
es-es	61.33%	42.52%	1086MB
it-it	71.92%	46.68%	170MB
fr-fr	53.47%	29.81%	487MB

The JIRS-based systems performed better than the Lucene-based ones in Spanish and French, whereas in Italian they obtained almost the same results. The difference in the CWS values obtained in both Spanish and French is consistent and weights in favour of JIRS. This prove that the quality of passages returned by JIRS for these two languages is considerably better.

We measured also the inter-agreement of the two systems, by counting the number of times that the two systems returned the same source document divided by the number of times that they returned the same answer (we called this measure *Agreement on Answer* or *AoA*), and when the answer was the right one (in this case we call it *AoRA*).

As it can be observed in Table 3, the best agreement has been obtained in Italian, as one would expect due to the smaller size of the collection.

4 Conclusions and Further Work

We obtained a slight improvement over the results of our 2005 system. This is consistent with the small amount of modifications introduced, principally because of the new rules defined for the 2006 CLEF QA task. The most interesting result is that JIRS demonstrated to be more effective for the QA task than a standard keyword-based search engine such as Lucene in two languages over three. Our further works on the QUASAR system will concern the implementation of a specialized strategy for definition questions, and probably a major revision of the Answer Extraction module.

Acknowledgments

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work. This paper is a revised version of the work titled “The UPV at QA@CLEF 2006” included in the CLEF 2006 Working Notes.

References

1. Aunimo, L., Kuuskoski, R., Makkonen, J.: Cross-language question answering at the university of helsinki. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, Springer, Heidelberg (2005)
2. Light, M., Mann, G.S., Riloff, E., Breck, E.: Analyses for elucidating current question answering technology. *Nat. Lang. Eng.* 7(4), 325–342 (2001)
3. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Multilingual question/answering: the DIOGENE system. In: The 10th Text REtrieval Conference (2001)
4. Moldovan, D., Pasca, M., Harabagiu, S., Surdeanu, M.: Performance issues and error analysis in an open-domain question answering system. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, New York, USA (2003)
5. Neumann, G., Sacaleanu, B.: Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, Springer, Heidelberg (2005)
6. Soriano, J.M.G., Buscaldi, D., Bisbal, E., Rosso, P., Sanchis, E.: Quasar: The question answering system of the universidad politécnica de valencia. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
7. Vallin, A., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peas, A., Rijke, M.: Overview of the clef 2005 multilingual question answering track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. Vicedo, J.L., Izquierdo, R., Llopis, F., Munoz, R.: Question answering in spanish. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, Springer, Heidelberg (2004)

Cross-Lingual Romanian to English Question Answering at CLEF 2006

Georgiana Pușcașu^{1,4,5}, Adrian Iftene¹, Ionuț Pistol¹, Diana Trandabăț^{1,3},
Dan Tufiș², Alin Ceașu², Dan Ștefănescu², Radu Ion²,
Iustin Dornescu^{1,3}, Alex Moruz^{1,3}, Dan Cristea^{1,3}

¹ UAIC: Faculty of Computer Science

"Alexandru Ioan Cuza" University, Romania

{adiftene, ipistol, dtrandabat, idornescu, mmoruz,
dcristea}@info.uaic.ro

² RACAI: Centre for Artificial Intelligence

Romanian Academy, Romania

{tufis, alceausu, danstef, radu}@racai.ro

³ Institute for Computer Science

Romanian Academy Iasi Branch

⁴ Research Group in Computational Linguistics

University of Wolverhampton, UK

{georgie}@wlv.ac.uk

⁵ Department of Software and Computing Systems

University of Alicante, Spain

Abstract. This paper describes the development of a Question Answering (QA) system and its evaluation results in the Romanian-English cross-lingual track organized as part of the CLEF¹ 2006 campaign. The development stages of the cross-lingual Question Answering system are described incrementally throughout the paper, at the same time pinpointing the problems that occurred and the way they were addressed. The system adheres to the classical architecture for QA systems, debuting with question processing followed, after term translation, by information retrieval and answer extraction. Besides the common QA difficulties, the track posed some specific problems, such as the lack of a reliable translation engine from Romanian into English, and the need to evaluate each module individually for a better insight into the system's failures.

1 Introduction

QA can be defined as the task which takes a question in natural language and produces one or more ranked answers from a collection of documents. The QA research area has emerged as a result of a monolingual English QA track being introduced at TREC². Multilingual, as well as mono-lingual QA tasks for languages other than English are addressed at a scientific level by the CLEF evaluation campaigns [8].

¹ Cross Lingual Evaluation Forum - <http://www.clef-campaign.org/>

² Text Retrieval and Evaluation Conference - <http://trec.nist.gov/>

This year our team of students, PhD students and researchers, have taken part for the first time in the QA@CLEF competition. As it was the first time the Romanian-English track was introduced at CLEF, most of the effort was directed towards the development of a fully functional cross-lingual QA system having Romanian as source language and English as target language, and less on fine-tuning the system to maximize the results.

Our system follows the generic architecture for QA systems [2] and contains a question analyzer, an information retrieval engine and an answer extraction module, as well as a cross-lingual QA specific module which translates the relevant question terms from Romanian to English.

The present paper describes the development and evaluation results of the cross-lingual QA system for the Romanian-English track at CLEF 2006. The remainder of the paper is structured as follows: Section 2 provides a description of the system and its embedded modules, while Section 3 presents the details of the submitted runs. Section 4 comprises an analysis of the results obtained; finally, in Section 5, conclusions are drawn and future directions of system development are considered.

2 System Description

2.1 System Overview

QA systems normally adhere to the pipeline architecture comprising three main stages: question analysis, paragraph retrieval and answer extraction [2]. The first stage is question analysis. The input to this stage is a natural language question and the output is one or more question representations to be used at subsequent stages. At this stage most systems identify the semantic type of the entity sought by the question, determine additional constraints on the answer entity, and extract keywords to be employed at the next stage. The passage retrieval stage is typically achieved by employing a conventional IR search engine to select a set of relevant candidate passages from the document collection. At the last stage, answer extraction and ranking, the representation of the question and the representation of the candidate answer-bearing passages are compared and a set of candidate answers is produced, ranked according to the likelihood of being the correct answer. Apart from the typical modules of a QA system, our system also includes a module that performs Romanian to English translation, thus achieving the cross-lingual transfer. Figure 1 illustrates the system architecture and functionality.

2.2 NLP Pre-processing

The questions are first morpho-syntactically pre-processed using the Romanian POS tagger developed at RACAI [9,3]. A pattern-based Named Entity (NE) Recognizer is then employed to identify Named Entities classified in one of the following classes: *Person*, *Location*, *Measure* and *Generic*. Similar pre-processing is applied to the English document collection with the same set of tools, but a different language model.

2.3 Question Analysis

This stage is mainly concerned with the identification of the semantic type of the entity sought by the question (expected answer type). In addition, it also provides the question

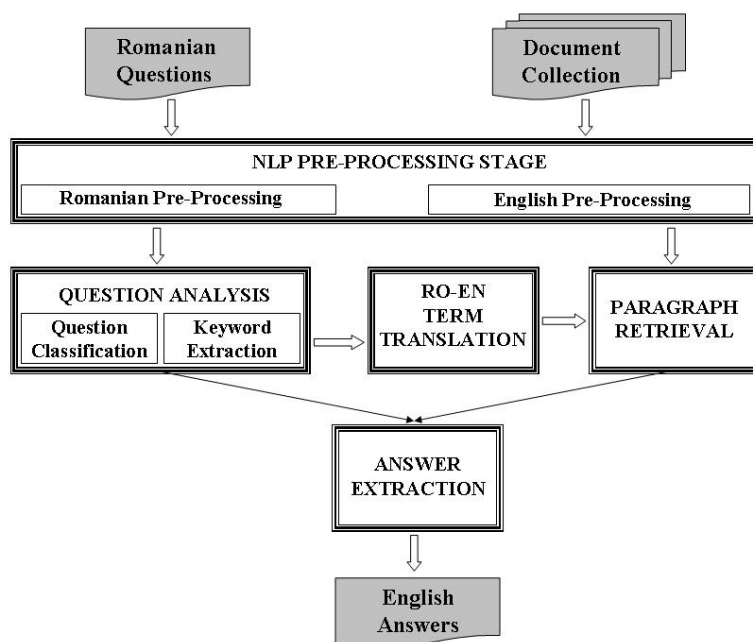


Fig. 1. System Architecture and Functionality

focus, the question type and the set of keywords relevant for the question. To achieve these goals, our question analyzer performs the following steps:

a) ***NP-chunking, Named Entity Extraction, Temporal Expression Identification***

The pre-processed set of questions serves as input. On the basis of the morpho-syntactic annotation provided by the Romanian POS-tagger, a rule-based shallow noun phrase identifier was implemented. The NE recognizer employed at the pre-processing stage provides us with the set of question NEs. Temporal expressions (TEs) are also identified using the adaptation to Romanian of the TE identifier and normalizer developed by [7].

b) ***Question Focus Identification***

The question focus is the word or word sequence that defines or disambiguates the question, in the sense that it pinpoints what the question is searching for or what it is about. The question focus is considered to be either the noun determined by the question stem (as in *What **country***) or the head noun of the first question NP if this NP comes before the question's main verb or if it follows the verb "to be".

c) ***Distinguishing the Expected Answer Type***

At this stage the category of the entity expected as an answer to the analyzed question is identified. Our system's answer type taxonomy distinguishes the following classes: PERSON, LOCATION, ORGANIZATION, TEMPORAL, NUMERIC, DEFINITION and GENERIC. The assignment of a class to an analyzed question is performed using the question stem and the type of the question focus. The question

focus type is detected using WordNet [1] sub-hierarchies specific to the categories PERSON / LOCATION / ORGANIZATION. We employ a pre-defined correspondence between each category and the ILI codes of WordNet root nodes heading category-specific noun sub-trees. These ILI codes guide the extraction of category specific noun lists from the Romanian WordNet [10,11]. In the case of ambiguous question stems (e.g. *What*), the resulted lists are searched for the head of the question focus, and the expected answer type is identified with the category of the corresponding list (for example, in the case of the question *In which city was Vladislav Listyev murdered?*, the question focus is *city*, noun found in the LOCATION list, therefore the associated expected answer type is LOCATION).

d) ***Inferring the Question Type***

This year, the QA@CLEF main task distinguishes among four question types: *factoid*, *definition*, *list* and *temporally restricted* questions. As temporal restrictions can constrain any type of question, we proceed by first detecting whether the question has the type *factoid*, *definition* or *list* and then test the existence of temporal restrictions. The question type is identified using two simple rules: if the expected answer type is DEFINITION, then obviously the question type is *definition*; if the question focus is a plural noun, then the question type is *list*, otherwise *factoid*. The temporal restrictions are identified using several patterns and the information provided by the TE identifier.

e) ***Keyword Set Generation***

The set of keywords is automatically generated by listing the important question terms in decreasing order of their relevance. Therefore, this set comprises: the question focus, the identified NEs and TEs, the remaining noun phrases, and all the non-auxiliary verbs present in the question. It is then passed on to the Term Translation module, in order to obtain English keywords for paragraph retrieval.

2.4 Term Translation

Term translation is achieved by means of WordNet, a resource available both for English and Romanian. The keywords extracted at the question analysis stage serve as input for the term translation module, so the translation process needs to deal with both noun phrases and verbs. Noun phrases are translated in a different manner than verbs. The NP words are translated individually, by first identifying the Romanian synsets containing the word, by reaching through ILI codes the corresponding English synsets, and by forming a set of all possible translations. This set is then reduced to a maximum of three candidates with empirical rules based on word frequency. If the word to be translated does not appear in the Romanian WordNet, as it is quite frequently the case, we search for it in other available dictionaries and preserve the first three translations. If still no translation is found, the word itself is considered as translation. After each word has been translated, rules are employed to translate the Romanian syntax to English syntax. In this manner, several translation equivalents are obtained for a given NP.

In the case of verbs, WordNet translation equivalents are extracted in the same manner as for nouns. Attempts to apply the frequency-based methodology employed for nouns failed due to the fact that certain general verbs are preferred to the correct translation equivalent. To this end, we decided to select the best translation equivalent by

considering both the frequency of the verb and of the nouns which appear in the subject and object positions. To achieve this, the verb-noun co-occurrence data described in [6] has been used to determine which of the verb translation equivalents co-occur more often with the translated nouns, and the most frequent three verbs were selected as translation. Despite the simplicity of the method, the accuracy of the translation improves dramatically over the row frequency.

2.5 Index Creation and Paragraph Retrieval

As described in section 2.2, the English corpus was initially preprocessed using tokenization, lemmatization, POS-tagging and NE recognition tools. In our runs, we employed an index/search engine based on the Lucene [5] search engine.

The document collection was indexed both at document, as well as paragraph level, using the lemmas of the content words and the NE classes (MEASURE, PERSON, LOCATION, etc). For a query like:

(HOW MANY, MEASURE)

(record book "record book") AND (sport athlete sportsman) AND (world worldwide) AND (man husband male) AND (score hit carry) AND (1995)

the search engine looks for a segment (document/paragraph) that contains a measure and words belonging to the set of translation equivalents. When no paragraphs are returned for a given query, two strategies are employed: either the granularity of the segments is increased from paragraphs to documents, or the query is reformulated using spelling variations for individual words and variable word distance for phrases.

2.6 Answer Extraction

Two answer extraction modules have been developed, one by UAIC and another one by RACAI. Both modules require as an input the expected answer type, the question focus, the set of keywords and the retrieved snippets together with their relevance scores returned by Lucene and their POS, lemma and NE information. The extraction process depends on whether the expected answer type is a Named Entity or not. When the answer type is a Named Entity, the answer extraction module identifies within each retrieved sentence Named Entities with the desired answer type. Therefore, in this case, the answer extraction process is greatly dependent on the results of the NE recognition module. When the answer type is not a Named Entity, the extraction process mainly relies on the recognition of the question focus, as in this case the syntactic answer patterns based on the focus are crucial. Both answer extraction modules employ the same pattern-based extractor to address DEFINITION questions.

a) *Answering Questions Asking about Named Entities*

When the question asks about a Named Entity such as MEASURE, PERSON, LOCATION, ORGANIZATION, DATE, the UAIC answer extractor looks for all the expressions tagged with the desired answer type. If such expressions exist, the answer is the closest one to the focus in terms of word distance, if the focus is present in the sentence, otherwise the first one occurring in the analyzed sentence. When

there is no Named Entity of the desired type, the search is generalized using synonyms of the focus extracted from WordNet.

The RACAI answer extractor computes scores for each snippet based on whether a snippet contains the focus, on the percentage of keywords or keyword synonyms present in the snippet, and on the snippet/document relevance scores provided by the search engine. The entities with the desired answer type are identified and added to the set of candidate answers. For each candidate answer, another score is computed on the basis of the snippet score, the distance to the question focus and to other keywords. The candidate answers having scores above a certain threshold are presented as final answers.

b) **Answering Questions Looking for *GENERIC* Answers**

When the expected answer type is not a Named Entity (the question has the *GENERIC* answer type), the UAIC answer extractor locates possible answers within candidate sentences by means of syntactic patterns. The syntactic patterns for identifying the answer include the focus NP and the answer NP, which can be connected by other elements such as comma, quotation marks, prepositions or even verbs. These syntactic patterns always include the question focus. Therefore, the focus needs to be determined by the question analysis module in order to enable the system to identify answers consisting of a noun or verb phrase.

The RACAI answer extractor employs also patterns to select answer candidates. These candidates are then ranked using a score similar to the one employed for answering questions asking about NEs.

c) **Answering *DEFINITION* Questions**

In the case of *DEFINITION* questions, the candidate paragraphs are matched against a set of patterns. Each possible definition is extracted and added to a set of candidate answers, together with a score revealing the reliability of the pattern it matched. The set of snippet NPs is also investigated to detect those NPs containing the term to be defined surrounded by other words (this operation is motivated by cases like *the Atlantis space shuttle*, where the correct definition for *Atlantis* is *space shuttle*). The selected NPs are added to the set of candidate answers with a lower score. Then, the set of candidate answers is ordered according to the score attached to each answer and to the number of other candidate answers it subsumes. The highest ranked candidate answers are presented as final answers.

3 Description of Submitted Runs

Our QA@CLEF 2006 participation involved three different runs, with the following detailed description:

– **RACAI**

This run was obtained by parsing and analyzing the questions, translating the keywords, retrieving relevant passages and finding the final answers using the RACAI answer extractor.

– **UAIC**

This run is also obtained by parsing and analyzing the questions, keyword translation, passage retrieval, and answer localization with the UAIC answer extractor.

– DIOGENE

Our third run, which unfortunately was not evaluated, was obtained by converting the results of our Question Analysis and Term Translation modules to the format required by the DIOGENE QA system [4], and then providing them as input to the DIOGENE IR and Answer Extraction modules.

Due to the high number of submitted runs having English as target language, only the UAIC and RACAI runs were evaluated. In the following, the RACAI run will be referred to as System 1, and the name System 2 will apply to the UAIC run.

4 Results and Analysis

4.1 General Evaluation Results

The official evaluation results of Systems 1 and 2 are presented in Figure 2. Each submitted answer was evaluated as UNKNOWN, CORRECT, UNSUPPORTED, INCORRECT and INEXACT.

Evaluation results for System 1		
Z	Unknown	400
R	Correct	35
U	Unsupported	13
W	Incorrect	184
X	Inexact	7
Total		639

Evaluation results for System 2		
Z	Unknown	543
R	Correct	22
U	Unsupported	4
W	Incorrect	191
X	Inexact	1
Total		761

Fig. 2. Results for the two evaluated runs

The high number of answers evaluated as UNKNOWN is due to the fact that we provided ten answers for almost all 200 questions, as ten was the maximum number of answers allowed. However, the final evaluation considered only the first answer for the majority of the questions (only in the case of *list* type questions the first three answers were evaluated). As a result of this, the answers ranked 2 to 10 were labeled as UNKNOWN indicating that no attempt was made to check their correctness. Internal evaluation has revealed that the correct answers lied in the first ten answers generated by the two systems for approximately 40% of the questions. This is a promising result, as it reveals that the answer extractors work, but a better answer ranking methodology is required.

Most of the UNSUPPORTED and INEXACT answers are caused by the lack of a perfect match between the answer string and the supporting snippet. Such errors were partly introduced at the pre-processing stage by removing certain characters that have caused an answer like *United States* to be marked as UNSUPPORTED because the related snippet contains *United_States*), even if it denotes the same entity as the gold answer. Also, sometimes the two runs provide a more generic, or more specific answer

than required, as it is the case for the gold answer *Shoemaker* contrasted to our answer *Carolyn Shoemaker*, marked as UNSUPPORTED. These errors could be corrected by improving the answer extractor with more specific rules as to the format of the required answer. Several UNSUPPORTED and INEXACT answers apply to *list* type questions, as both runs provided one answer per line, even if the question is of the type *list*. Most of the correct answers can be found in the 10 answers provided, but not on the same line. These errors will be solved with a better classification of the returned answers and by grouping them according to the characteristics of the required list.

4.2 Detailed Comparison Between the Evaluated Runs

The two runs chosen for the official evaluation are the two completely designed and implemented by the two Romanian groups involved, UAIC and RACAI. They were selected from the set of three submitted not because they provide better scores, but because their source code and the knowledge of their original implementation can be easily accessed. As a result of this, we expect to significantly improve them in the future.

The performance of the two systems is compared below for each answer type separately, as the answer extraction methodology differs from one answer type to another.

In the case of the PERSON answer type, System 1 has correctly answered a higher number of *factoid* and *temporally restricted* questions, while System 2 has only detected correct answers for a small number of *temporally restricted* questions. There are many cases when the correct answer was retrieved, but not returned as the first answer. The improvement of both answer extraction modules and of the NER module employed at pre-processing need to be seriously addressed for a future CLEF participation.

System 1 has also performed better than System 2 for questions of LOCATION answer type. The best results were obtained for *list* questions asking about LOCATIONS (30% of the LOCATION *list* questions were correctly answered by System 1, while in the case of System 2 most answers were evaluated as UNSUPPORTED). For a number of questions asking for LOCATIONS, due to changes introduced by snippet pre-processing, answers correctly detected by System 2 were marked as UNSUPPORTED because the indicated snippet contained an “_” instead of a whitespace. Therefore, we need to ensure that our answer extractor preserves the answer format according to the original snippet.

For questions asking about ORGANIZATIONS, System 2 achieved better results than System 1, but still quite low. Both systems failed, mainly because the NE Recognizer did not identify Named Entities of type ORGANIZATION. Therefore, when the search engine was looking for an entity of this type, it was obviously impossible for it to retrieve any snippets containing such entities.

In the case of the DATE expected answer type, the questions asking about a DATE and being at the same time *temporally restricted* proved to be the most difficult questions to process. Both systems managed to find correct answers only for the *factoid* DATE questions, but no correct answer was found for *temporally restricted* DATE questions. In the case of System 1, most of the answers given to the *temporally restricted* DATE questions were considered UNSUPPORTED by the provided snippet. A large number of UNSUPPORTED answers for System 1 were due to truncating the extracted paragraphs to comply with the 500-byte limit, and leaving the answer in the truncated

part. In any case, the improvement of the answer extractor is required with respect to the identification of the correct extent of the required answer. System 2 answered correctly less *factoid* DATE questions in comparison with System 1. In the future, a temporal pre-annotation of the corpus will be performed, in order to facilitate the search for TEs and to be able to provide an answer at the granularity required by the question.

For the MEASURE answer type, the answer extractors mainly failed due to errors in the annotation introduced at the pre-processing stage. For a number of questions, the snippets were correctly selected (the gold answers lie in these snippets), but wrong answers were extracted due to wrong annotation.

All the questions that did not fit in the previous categories (PERSON, LOCATION, ORGANIZATION, DATE and MEASURE) were classified as GENERIC. For the GENERIC answer type, System 1 had significantly more correct *list* and *definition* answers when compared to System 2.

As a general overview, it has been noticed that System 1 outperformed System 2 for most answer types. Bearing in mind future competitions, the beneficial features of each system will be merged, at the same time improving each developed module for an increased performance.

5 Conclusions

This paper describes the development stages of a cross-lingual Romanian-English QA system, as well as its participation in the QA@CLEF campaign. We have developed a basic QA system that is able to retrieve answers to Romanian questions in a collection of English documents.

Adhering to the generic QA system architecture, the system implements the three essential stages (question analysis, information retrieval and answer extraction), as well as a cross-lingual QA specific module which translates the relevant question terms from Romanian to English.

The participation in the QA@CLEF competition included three runs. Two runs were obtained by analyzing the questions, translating the keywords, retrieving relevant passages and finding the final answers using two different answer extractors. The third run, which unfortunately was not evaluated, was provided by the DIOGENE QA system on the basis of the output given by our Question Analysis and Term Translation modules.

The results are poor for both evaluated runs, but we declare ourselves satisfied with the fact that we have managed to develop a fully functional cross-lingual QA system and that we have learned several important lessons for future participation.

A detailed result analysis has revealed a number of major system improvement directions. The term translation module, as a key stage for the performance of any cross-lingual QA system, is the first improvement target. However, the answer extraction module is the one that requires most of our attention in order to increase its accuracy. An improved answer ranking method for the candidate answers will also constitute a priority.

The development of the first Romanian-English cross-lingual QA system has provided us with valuable experience, at the same time setting the scene and motivating us for subsequent CLEF participations.

Acknowledgements

The authors would like to thank Constantin Orăsan and the following members of the UAIC team: Corina Forăscu, Maria Husarciuc, Mădălina Ioniță, Ana Masalagiu, Marius Răschip, for their help and support at different stages of system development.

Special acknowledgements are addressed to Milen Kouylekov and Bernardo Magnini for providing us with the DIOGENE run, by processing the output of our question processor and term translation module with the IR and answer extraction modules included in DIOGENE.

The research described in this paper has been partially funded by the Romanian Ministry of Education and Research under the CEEX 29 ROTEL project and by INTAS under the RoITech (ref. 05-104-7633) project.

References

1. Fellbaum, C.: WordNet: An Electronic Lexical Database. The MIT Press, Cambridge (1998)
2. Harabagiu, S., Moldovan, D.: Question Answering. *Oxford Handbook of Computational Linguistics*, 560–582 (2003)
3. Ion, R.: Automatic semantic disambiguation methods. Applications for English and Romanian (2006)
4. Kouylekov, M., Magnini, B., Negri, M., Tanev, H.: ITC-irst at TREC-2003: the DIOGENE QA system. In: *Proceedings of the Twelfth Text Retrieval Conference (TREC-12)* (2003)
5. LUCENE, <http://lucene.apache.org/java/docs/>
6. Pekar, V., Krkoska, M., Staab, S.: Feature weighting for cooccurrence-based classification of words. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)* (2004)
7. Puscasu, G.: A Framework for Temporal Resolution. In: *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC2004)* (2004)
8. QA@CLEF (2006), <http://clef-qa.itc.it/CLEF-2006.html>
9. Tufis, D.: Tagging with Combined Language Models and Large Tagsets. In: *Proceedings of the TELRI International Seminar on Text Corpora and Multilingual Lexicography* (1999)
10. Tufis, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal on Information Science and Technology*. Special Issue on BalkaNet (2004)
11. Tufis, D., Barbu Mititelu, V., Ceausu, A., Bozianu, L., Mihaila, C., Manu Magda, M.: New developments of the Romanian WordNet. In: *Proceedings of the Workshop on Resources and Tools for Romanian NLP* (2006)

Finding Answers in the C dipe System by Extracting and Applying Linguistic Patterns

Romaric Besan on, Mehdi Embarek, and Olivier Ferret

CEA-LIST

LIC2M (Multilingual Multimedia Knowledge Engineering Laboratory)

B.P.6 - F92265 Fontenay-aux-Roses Cedex, France

{besanconr,embarekm,ferreto}@zoe.cea.fr

Abstract. The simple techniques used in the C dipe question answering system developed by the CEA-LIST/LIC2M did not perform well on definition questions in past CLEF-QA campaigns. We present in this article a new module for this QA system dedicated to this type of questions. This module is based on the automatic learning of lexico-syntactic patterns from examples. These patterns are then used for the extraction of short answers for definition questions. This technique has been experimented in the French monolingual track of the CLEF-QA 2006 evaluation, and results obtained have shown an improvement on this type of questions, compared to our previous participation.

1 Introduction

The C dipe question-answering system developed by the CEA-LIST/LIC2M was primarily a baseline system that is progressively extended by adding new modules in order to both enlarge its capabilities and improve its results. This incremental approach allows to carefully test the impact and the interest of each added module. The first version of the C dipe system [1] was a passage-based system that has been developed for the French EQUER evaluation campaign [2]. It has been extended for CLEF-QA 2005 [3] to extract short answers from passages. Our results in this evaluation showed that our strategy for processing factoid questions was reasonably successful but that our few heuristics for answering definition questions were not sufficient. Hence, for our participation to the French monolingual track of the CLEF-QA 2006 campaign, we developed a new module in the C dipe system that is designed to deal with definition questions. This module is based on the automatic learning of patterns to extract short answers for definition questions.

We present in section 2 an overview of the C dipe system. We then focus in section 3 on the techniques used in the new module added for definition questions. Finally, an evaluation and discussion of our results is presented in section 4.

2 Overview of the $\mathcal{E}$ dipe System

We present in this section the main features of the $\mathcal{E}$ dipe system. More details can be found in the report of the previous campaign [3]. Figure 1 illustrates the architecture of the $\mathcal{E}$ dipe system. This architecture is a standard pipeline archi-

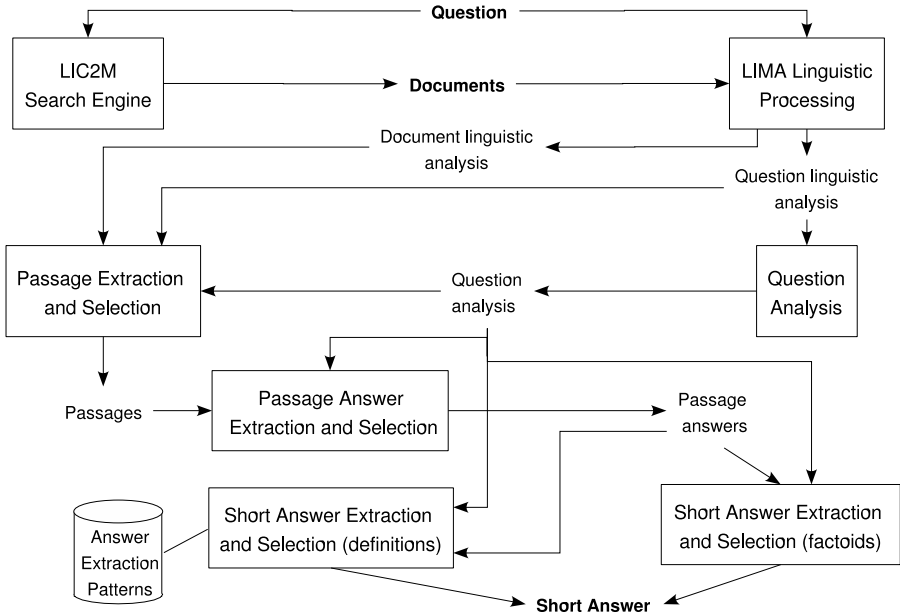


Fig. 1. Architecture of the $\mathcal{E}$ dipe system

ture: the question is first submitted to a search engine to retrieve a restricted set of documents. The linguistic processing provided by the LIMA (LIC2m Multilingual Analyzer) analyzer [4] is used to normalize words and extract named entities from both question and documents. A deeper analysis of the question is performed to identify the expected type of the answer and the focus of the question. A two-step gisting process is applied to the documents retrieved by the search engine to locate where answers are more likely to be found: passages are first delimited using the density of the question words in documents, which is a basic criterion but is not computationally expensive; passage answers with a maximal size of 250 characters are then extracted from these passages by computing a score in a sliding fixed-size window. Finally, short answers¹ are extracted from the passage answers depending on their expected type.

More precisely, the $\mathcal{E}$ dipe system is composed of the following modules:

¹ “Short answer” refers here to the exact answer and not to a fixed-size answer.

Search Engine. As in CLEF-QA 2005, we used the LIC2M search engine, that was already evaluated through the Small Multilingual Track of CLEF in 2003 and 2004 [5,6]. Each question is submitted to the search engine without any pre-processing as the LIC2M search engine applies the LIMA linguistic processing to its queries. The only difference with CLEF-QA 2005 concerns the number of retrieved documents. In CLEF-QA 2005, a predefined number of documents (equal to 25) was retrieved. But the LIC2M search engine is concept-based and returns classes of documents that share the same query concepts (*i.e.* multi-terms and named entities): classes are ranked, but documents inside classes are considered equivalent and are not ranked. Hence, retrieving a given number of documents implies an arbitrary cut inside a class. We tried to avoid this arbitrary selection by using a variable number of documents (between 20 and 50) and applying the following algorithm:

```

selectedDocuments  $\leftarrow \{\}$ 
i  $\leftarrow 1$ 
while card(selectedDocuments) < 20  $\wedge$  i  $\leq$  card(classes) do
    currentClass  $\leftarrow$  classes[i]
    i  $\leftarrow$  i + 1
    if card(selectedDocuments) + card(currentClass)  $\leq$  50
        then
            selectedDocuments  $\leftarrow$  selectedDocuments  $\cup$  currentClass
        else
            randSelDocsNb = 50 - card(selectedDocuments)
            selectedDocuments  $\leftarrow$  selectedDocuments  $\cup$ 
                random(currentClass, randSelDocsNb)
    fi
fi

```

where *random*(*s*, *n*) is a function that randomly selects *n* elements from a set *s*. For CLEF-QA 2006, an average number of 33 documents by question was selected by this algorithm.

Linguistic Processing. The linguistic processing of questions and documents is performed by the LIMA linguistic analyzer. Only a subset of its features² is used to normalize words (using a POS-tagger and lemmatizer), identify content words and extract named entities. The named entities recognized by LIMA are restricted to the MUC named entities [7], that is to say *persons*, *locations*, *organizations*, *dates* and *times* and *numerical measures*, plus *events* and *products*.

Question Analysis. In (Edipe, the question analysis module performs three different tasks:

- *identification of the expected type of the answer*: the result of this task determines the strategy applied by (Edipe for extracting answers. If the expected type of the answer corresponds to a type of named entities that can be recognized by LIMA, (Edipe searches in the selected document

² The LIMA analyzer can also perform term extraction or syntactic analysis but these capabilities are not exploited in (Edipe yet.

passages for the named entity of that type whose context is the most compatible with the question. Otherwise, it assumes that the question is a definition question and applies specific linguistic patterns to extract possible answers (see Section 3);

- *identification of the focus of the question*: the focus of a question is defined as the part of the question that is expected to be present close to the answer. This task is part of the new module developed for definition question³ and is achieved using a specific set of patterns, implemented as finite-state automata. Here are two examples of such patterns:

```
[Qu']::[est] [ce] [@Que] [$L_DET] *{1-30} [\?]:FOCUS:
[Qui]::[être$L_V] *{1-30} [\?]:FOCUS
```

The first rule identifies *Atlantis* as the focus of the definition question “*Qu’est-ce que l’Atlantis ?*” (*What is Atlantis?*) while the second rule extracts *Hugo Chavez* as the focus of the question “*Qui est Hugo Chavez ?*” (*Who is Hugo Chavez?*);

- *selection and weighting of the significant words of the question*: the significant words of the question are its content words and their weight is obtained by looking up their normalized information in a reference corpus to evaluate their specificity degree.

Passage Extraction and Selection. This module first delimits candidate passages by detecting the document areas with the highest density of question words. Then, a score is computed for each delimited passage depending on the number and the significance of the words of the question it contains. Candidate passages are ranked according to this score and passages whose score is lower than a predefined threshold are discarded.

Passage Answer Extraction and Selection. A passage answer is extracted from each selected passage by sliding a window over the passage and computing a score at each position of the window according to its content and the expected type of the answer. The size of this window is equal to the size of the passage answers to extract (250 characters in our case). The position of the window is anchored on content words for definition questions and on named entities that are compatible with the expected answer type for factoid questions. A predefined number of passage answers is selected according to their score.

Short Answer Extraction and Selection. When the expected type of the question is a named entity, each passage answer is centered on a named entity of that type. Hence, the extracted short answer is directly that named entity. Its score is equal to the score of the passage answer. The extraction of short answers for definition questions is presented in details in the next section. As for passage answers, each short answer is given a score and all short answers are ranked according to this score.

³ It is currently performed only for definition questions but from a more general point of view, it could also be useful for improving the processing of factoid questions.

3 Answering Definition Questions

As mentioned before, the main improvement of the Edipe system for CLEF-QA 2006 concerns its ability to answer questions whose expected answer is not a named entity, and more specifically definition questions such as *What is X?* or *Who is X?*. *What/Who* questions have proved to be difficult both because they are generally short and the search of an answer cannot be focused by a specific type of elements such as a named entity. Hence, trying to answer this kind of questions with a basic approach leads to poor results (as it was illustrated by Edipe's results at CLEF-QA 2005).

Most of the question answering systems rely on a set of handmade linguistic patterns to extract answers for that kind of questions from selected sentences [8]. Some work has been done to learn such patterns from examples, following some work in the Information Extraction field. One of the first attempt was from Ravichandran and Hovy [9], who proposed a strategy that combines the use of the Web and suffix trees. Mining the Web for learning such patterns is also the solution adopted by [10]. Jousse et al. [11] tested several machine learning algorithms for extracting patterns and Cui et al. [12] proposed a new algorithm for learning probabilistic lexico-syntactic patterns, also called *soft patterns*.

Works such as [9] or [11] have proved that building a set of question-answer examples is quite easy, especially from the Web. This is why we chose to rely on lexico-syntactic patterns learnt from examples for answering definition questions. Moreover, this approach appears to be both more flexible and less costly when a question answering system must be extended to new domains.

3.1 Learning of Definitional Patterns

The algorithm we used to learn linguistic patterns for extracting answers to definition questions is an extension of the Ravichandran and Hovy's algorithm [9]. This extension, proposed by Ravichandran in [13], allows to learn multilevel patterns instead of surface patterns: multilevel patterns can refer to different levels of linguistic information. This kind of patterns have been used in various applications to extract different kind of information: semantic relations to populate knowledge bases in [14] and [15], answers in question answering systems or factual relations between entities in information extraction.

In our case, the induction of patterns is done from a set of example answers to definition questions. The basic element of a pattern can be the surface form of a word, its part of speech (POS) or its lemma. These three levels of information are obtained using the LIMA linguistic analyzer. More precisely, the overall procedure for building up a base of patterns dedicated to the extraction of definitional answer is the following:

1. Building a corpus of example answer sentences. Unlike [9] or [10], our corpus of example answers was not built from the Web but came from the results of the EQUER evaluation and from the previous CLEF-QA evaluations. For each definition question of these evaluations, all the sentences containing a correct answer to the question were extracted.

2. Application of the LIMA linguistic analyzer to all the answer sentences to get the three levels of linguistic information.
3. Abstraction of answer sentences. This abstraction consists in replacing in each answer sentence the focus of the question by the tag *<focus>* and the short answer to the question by the tag *<answer>*.
4. Application of the multilevel pattern-learning algorithm between each pair of sentences (see below).
5. Selection of the top P patterns on the basis of their frequency.

The multilevel algorithm for the induction of patterns is taken from [13]. It is composed of two steps: the first one consists in calculating the minimal edit distance between the two sentences to generalize⁴; the second one extracts the most specific multilevel pattern that generalizes the two sentences. Some of the obtained alignments are then completed by adding two wildcard operators to take into account the presence of words between aligned words: *(*s*)* stands for 0 or 1 word while *(*g*)* stands for exactly 1 word.

Here are some examples of the definitional patterns induced by this algorithm⁵:

```

<answer> ( <focus>
<focus> ( <answer>
<focus> , le <answer>
<focus> , un <answer>
<focus> être L_DET_ARTICLE_INDEF <answer>
<focus> , (*g*) <answer>

```

3.2 Application of Patterns to Extract Answers

The definition patterns resulting from the learning method described in the previous section were integrated into the Cedipe system by applying them after the selection of passage answers. The exact procedure is the following:

1. Instantiation of definitional patterns: the *<focus>* tags in patterns is replaced by the focus of the question, as identified by the dedicated rules of the question analysis module (see Section 2).
2. Application of the LIMA linguistic analyzer: each passage answer is analyzed to get the three levels of linguistic information possibly used in patterns.
3. Extraction of short answers: each instantiated pattern is aligned with the passage answer. This alignment starts from the focus of the window and is checked word by word until the *<answer>* tag is reached in the pattern. If this alignment succeeds, the noun phrase in the passage answer that corresponds to the *<answer>* tag in the pattern is kept as a possible short answer.
4. Selection of the top A short answers on the basis of their frequency.

⁴ The edit distance between two sentences is equal to the number of edit operations (insertion, deletion and substitution) that are necessary to transform one sentence into the other one.

⁵ Patterns' lexical items are translated as: "*le*"=*the*, "*un*"=*a-an*, "*être*"=*to be*.

The set of short answers extracted by using definition patterns are ranked according to their number of matching patterns. The short answer with the highest value is returned as the answer to the considered question.

4 Evaluation

4.1 Results

Only one run was submitted in the CLEF-QA 2006 evaluation for the Edipe system. For the 187 factoid (F), definition (D) and temporally restricted (T) questions, Edipe returned 30 right answers, 3 unsupported answers and 6 inexact answers, which gives an overall accuracy of 16%. Moreover, the detection of a lack of answer by Edipe was right for only one question out of four. For the 10 list questions, 7 right answers were returned, which corresponds to an average precision of 0.1633 (see [16] for explanations about evaluation measures).

Table 1. Comparison of the distributions of Edipe’s right answers at CLEF-QA 2005 and CLEF-QA 2006

	Factoid (F+T)		Definition (D)	
	# right answers	accuracy	# right answers	accuracy
CLEF-QA 2005	28	18.7	0	0.0
CLEF-QA 2006	15	10.3	15	36.6

At a quick glance, the results of Edipe at CLEF-QA 2006 seem to be comparable, with a slight improvement to its results at CLEF-QA 2005, where its overall accuracy was equal to 14% with 28 right answers for the F, D and T questions. However, Table 1 shows that the distributions of right answers are quite different for the two evaluations. The introduction of definitional patterns for processing definition questions leads to a very significant improvement. On the contrary, results for factoid questions significantly decrease, with no relation to the previous fact since they were processed with the CLEF-QA 2005 version of Edipe.

4.2 Discussion

The improvement for definition questions was expected since a specific module was developed for that kind of questions but there is no clear explanation of the decrease of results for factoid questions, except that they were perhaps more difficult than CLEF-QA 2005 factoid questions. Moreover, Table 2 shows that the question analysis module is not responsible of this decrease of Edipe’s results for factoid questions since its accuracy for CLEF-QA 2006 questions is higher than for CLEF-QA 2005 questions. It should also be noted that the focus was correctly identified for all the definition questions.

Table 2. Results of the question analysis module for CLEF-QA 2006 and comparison with CLEF-QA 2005

question type	# questions	# incorrect types	accuracy (2006)	acc. (2005)
Factoid (F+T)	146	9	93.8	86.0
Definition (D)	41	4	90.2	100

Table 3. Detailed results of $\mathcal{E}$ dipe for the French monolingual track, both for CLEF-QA 2006 and CLEF-QA 2005

# answers / question	exact answer				passage answer			
	MMR		# right answers		MMR		# right answers	
	2005	2006	2005	2006	2005	2006	2005	2006
1	0.140	0.162	28	32	0.170	0.162	34	32
2	0.147	0.183	31	40	0.182	0.185	39	41
3	0.151	0.195 ⁶	33	47	0.193	0.197	45	48
4	0.152	0.195	34	47	0.194	0.198	46	49
5	0.152	0.195	34	47	0.197	0.201	49	52
10	0.154	0.195	37	47	0.203	0.203	59	54

As for CLEF-QA 2005, we exploited the assessed runs of all the participants to CLEF-QA 2006 with French as target language to build a gold standard for automatic investigations. This is an incomplete reference as answers were found for 185 questions among the 197 questions used for the official evaluation in [16] but this approach proved to be reliable. Each reference answer was transformed into pattern through case and separator normalization. The automatic comparison of a reference answer with an answer produced by $\mathcal{E}$ dipe was done by verifying both the matching with its pattern and the document in which it was found.

This gold standard was first used for evaluating the document retrieval module of $\mathcal{E}$ dipe. More precisely, the LIC2M search engine retrieved at least one relevant document for 135 questions, which represents 68.5% of questions. This is quite similar to the results for CLEF-QA 2005 (132 questions), which means that taking into account the specificities of the LIC2M search engine doesn't have a significant impact on $\mathcal{E}$ dipe's results. Hence, other solutions have to be found to increase this percentage as we could expect that a relevant document can be found for at least 85% of questions [17].

However, our gold standard mainly aims at investigating the capabilities of $\mathcal{E}$ dipe for answer extraction. Table 3 shows the results of $\mathcal{E}$ dipe for two kinds of answers (exact or short vs. passage answers) at several ranks (between 1 and 10) and compares them to its results at CLEF-QA 2005. First, it must be noted that with our gold standard, $\mathcal{E}$ dipe has 32 right short answers at first rank for

⁶ MRR at rank 3 is equal to 0.1907 in [16], which shows that our gold standard is reliable.

CLEF-QA 2006 instead of 30 in the official results [16] because this reference also takes into account list questions. For these questions, we consider that a question is answered if at least one item of the list is found, which facilitates comparisons with CLEF-QA 2005 results.

The overall comparison of *Œdipe*'s results for CLEF-QA 2005 with those for CLEF-QA 2006 shows the global improvement of our system about answer extraction. The number of right exact answers is the most evident sign of this improvement: a maximal value of 47 answers is reached at rank 3 for CLEF-QA 2006 whereas the maximal value for CLEF-QA 2005 is equal to 37 answers and is reached only at rank 10. The second sign of this improvement is given by the comparison of the results for exact and passage answers. While for CLEF-QA 2005, there is a significant difference between the results for the two kinds of answers, this difference is small (except for higher ranks) for CLEF-QA 2006. This means that when a passage that contains an answer is found, an exact answer is likely to be found. But this also means that one of the main bottlenecks of *Œdipe* is now its ability to find relevant passages.

5 Conclusion

In this article, we have given an overview of the version of the *Œdipe* system that participated to the French monolingual track of the CLEF-QA 2006 evaluation and we more particularly detailed its new aspects with regards to the CLEF-QA 2005 version, that is to say the learning from examples of lexico-syntactic patterns and their application to extract short answers for definition questions. These new aspects proved to be interesting but there is still room for improvements. The first of them is certainly to integrate the syntactic analysis of LIMA in *Œdipe*, which would particularly enable to take each noun phrase as a whole and make extraction patterns more general. The second main improvement is to extend the use of extraction patterns to the processing of factoid questions. The results of *Œdipe* for that kind of questions are significantly lower than its CLEF-QA 2005 results and an explanation of this fact still have to be found. However, it is clear that the use of such patterns should be an interesting way to focus the search of a named entity answer in passages.

References

1. Balvet, A., Embarek, M., Ferret, O.: et question-réponse: le système Oedipe. In: 12^{ème} Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005), Dourdan, France, pp. 77–80 (2005)
2. Ayache, C.: Campagne EVALDA/EQUER : Evaluation en question-réponse, rapport final de la campagne EVALDA/EQUER. Technical report, ELDA (2005)
3. Besançon, R., Embarek, M., Ferret, O.: The *Œdipe* System at CLEF-QA 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)

4. Besançon, R., de Chalendar, G.: L'analyseur syntaxique de LIMA dans la campagne d'évaluation EASY. In: 12^{ème} Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005), Dourdan, France, pp. 21–24 (2005)
5. Besançon, R., de Chalendar, G., Ferret, O., Fluhr, C., Mesnard, O., Naets, H.: Concept-based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 174–184. Springer, Heidelberg (2004)
6. Besançon, R., Embarek, M., Ferret, O.: Integrating New Languages in a Multilingual Search System Based on a Deep Linguistic Analysis. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 83–89. Springer, Heidelberg (2005)
7. Grishman, R., Sundheim, B.: Design of the MUC6 evaluation. In: MUC-6 (Message Understanding Conferences, Columbia, MD, Morgan Kaufmann Publisher, San Francisco (1995)
8. Soubotin, M.M., Soubotin, S.M.: Patterns of potential answer expressions as clues to the right answer. In: NIST, (ed.) TREC-10 Conference, Gaithersburg, MD, pp. 175–182 (2001)
9. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, pp. 41–47 (2002)
10. Du, Y., Huang, X., Li, X., Wu, L.: A novel pattern learning method for open domain question answering. In: Su, K.-Y., Tsujii, J., Lee, J.-H., Kwong, O.Y. (eds.) IJCNLP 2004. LNCS (LNAI), vol. 3248, pp. 81–89. Springer, Heidelberg (2005)
11. Jousse, F., Tellier, I., Tommasi, M., Marty, P.: Learning to extract answers in question answering: Experimental studies. In: CORIA, pp. 85–100 (2005)
12. Cui, H., Kan, M.Y., Chua, T.S.: Generic soft pattern models for definitional question answering. In: 28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR 2005), Salvador, Brazil (2005)
13. Ravichandran, D.: Terascale Knowledge Acquisition. PhD thesis, University of Southern California (2005)
14. Pantel, P., Ravichandran, D., Hovy, E.: Towards terascale knowledge acquisition. In: International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, pp. 771–777 (2004)
15. Embarek, M., Ferret, O.: Extraction de relations sémantiques à partir de textes dans le domaine médical. In: JOBIM 2006, Bordeaux, France (2006)
16. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the clef 2006 multilingual question answering track. In: Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop (CLEF 2006), Alicante, Spain (2006)
17. Grau, B., Ligozat, A.L., Robba, I., Vilnat, A., Bagur, M., Séjourné, K.: The bilingual system musclef at qa@clef 2006. In: Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop (CLEF 2006), Alicante, Spain (2006)

Question Answering Beyond CLEF Document Collections

Luís Costa

Linguateca / SINTEF ICT,
Pb 124 Blindern, 0314 Oslo, Norway
luis.costa@sintef.no

Abstract. Esfinge is a general domain Portuguese question answering system, which uses the information available in the Web as an additional resource when searching for answers. The experiments described in this paper show that the Web helped more with this year's questions than in previous years and that using the Web to support answers also enables the system to answer correctly or at least to find relevant documents for more questions. In this third participation in CLEF, the main goals were (i) to check whether a database of word co-occurrences could improve the answer scoring algorithm and (ii) to get higher benefits from the use of a named entity recognizer that was used last year in sub-optimal conditions. 2006 results were slightly better than last year's, even though the question set included more definitions of type *Que é X?* (What is X?), with which the system had the worst results in previous participations.

Keywords: Question answering, Named entity recognition, Measurement, Performance, Experimentation.

1 Introduction

Esfinge is a question answering system developed for Portuguese inspired on the architecture proposed by Eric Brill in [1]. Brill argues that it is possible to get interesting results by applying simple techniques to large quantities of data, using redundancy to compensate for lack of detailed analysis. The Web in Portuguese (see [2] for a characterization of it) can be an interesting resource for such architecture.

Esfinge participated in CLEF both in 2004 and 2005 editions demonstrating the applicability of this approach to Portuguese [3, 4]. The main goal for the development of the system this year was to refactor and improve the code developed in previous years with a modular approach to make the work useful for other people interested in question answering, and to make future developments easier. In addition to that we also managed in this participation in CLEF 2006 to test the use of a new resource (a database of word co-occurrences) [5], benefit more from use of a named entity recognizer [6] and make experiments to measure the gains in using the Web.

This paper describes the architecture of Esfinge at QA@CLEF 2006, presents and discusses the results of official and non-official runs, describes an experiment where the Web was used to support answers and concludes with some remarks about the evolution of CLEF and Esfinge.

2 Esfinge 2006

Figure 1 gives a general overview of the architecture of Esfinge this year:

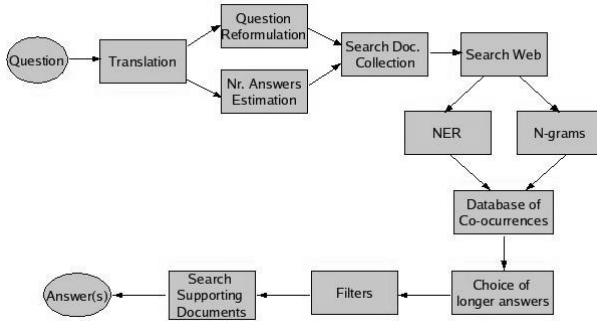


Fig. 1. Modules used in Esfinge

If the questions are not in Portuguese, they are translated to Portuguese. Then, Esfinge estimates the number of answers for the question. It proceeds with the Question Reformulation Module which transforms the question into patterns of plausible answers. These patterns are then searched in the document collection using the Search Document Collection module.

If the patterns are not found in the document collection, the system returns the answer NIL (no answer found) and stops. Otherwise, it proceeds by searching the same patterns in the Web. Then, the texts retrieved from the Web and the document collection are analyzed using the named entity recognizer (NER) system SIEMÈS [6] and an n-grams module in order to obtain candidate answers. The candidate answers are then ranked according to their frequency, length and the score of the passage from where they were retrieved. This ranking is in turn adjusted using the database of co-occurrences BACO [5] and the candidate answers (by ranking order) are analyzed in order to check whether they pass a set of filters and whether there is a document in the collection which supports them (this is necessary because Esfinge uses also text from the Web).

From the moment when Esfinge finds the estimated number of answers for the question, it checks only candidate answers that include one of the previously found answers. It will replace one of the original answers if the new one includes the original answer, passes the filters and has documents in the collection that can support it. For example, for the question *O que é a Generalitat?* (What is the Generalitat?), Esfinge found the answer *Catalunha* first, but found afterwards the candidate answer *governo da Catalunha* (Government of Catalonia) that includes the first and satisfies all the other criteria. Therefore the answer returned by the system was *governo da Catalunha*.

In the remainder of this section we will focus on the new modules or modules where there were some changes from last year's participation. For a complete description of Esfinge, please check [4].

2.1 Translation

This module was used for the EN-PT run (questions in English, answers in Portuguese). The questions (except the ones of type *Where is X*) are translated with the module `Lingua::PT::Translate` freely available at CPAN which provides an easy interface to the Altavista Babelfish translating tool. Questions of type *Where is X* are treated in a different way because the translation tool translates them to *Onde está X* (*is* can be translated as *está*, *é* or *fica* in Portuguese). Therefore this module translates these questions to *Onde fica X* which is a more natural translation.

2.2 Number of Answers Estimation

The motivation for this module arose from the inclusion of list questions in the question set this year. To estimate the number of answers required by a question, Esfinge submits it to the PALAVRAS parser [7], and:

- If the noun phrase (NP) is singular Esfinge tries to find exactly one answer (e.g. *Quem era o pai de Isabel II?* / *Who is Elizabeth The Second's father?*).
- If the NP is plural and the parser finds a numeral in the question (e.g. *Quais são as três repúblicas bálticas?* / *Which are the three Baltic states?*), the system will return this number of answers (three in this case).
- If the NP is plural and the parser does not find a numeral in the question (e.g. *Quais são as repúblicas bálticas?* / *Which are the Baltic states?*), Esfinge returns five answers (default).

This strategy is quite naive however, and after analyzing the question set we quickly realize that for some questions with a singular NP, it is more useful to obtain several answers (e.g. *Quem é Stephen Hawking?* / *Who is Stephen Hawking?* , *O que é o hapkido?* / *What is Hapkido?*, *Diga um escritor chinês.* / *Mention a Chinese writer.*).

2.3 Web Search

The patterns obtained in the previous modules (which have an associated score that indicate how likely they are to retrieve relevant passages) are then searched in the Web. This year in addition to Google's search API we also used Yahoo's. Esfinge stores the first 50 snippets (excluding, however, the snippets retrieved from addresses containing words like *blog* or *humor*) retrieved by each of the search engines and the scores of the patterns used to retrieve them $\{(\text{Snippet}_1, \text{Score}_1), (\text{Snippet}_2, \text{Score}_2) \dots (\text{Snippet}_n, \text{Score}_n)\}$.

2.4 Named Entity Recognition and N-Gram Harvesting

Two techniques are used to extract answers from the relevant passages retrieved from the Web and the document collection: named entity recognition and n-gram harvesting.

As reported in [8], 88,5% of the questions in CLEF 2004 referred explicitly named entities (NE) and 60,5% of the answers were simply NEs or included NEs. The NER system SIEMÊS [6] is used in Esfinge for questions which imply answers of type *Place*, *People*, *Quantity*, *Date*, *Height*, *Duration*, *Area*, *Organization* and *Distance*.

Esfinge uses pattern matching to check whether it is possible to infer the type of answer for a given question. For example, questions starting with *Onde* (Where) imply an answer of type *Place*, questions starting with *Quando* (When) imply an answer of type *Date*, questions starting with *Quantos* (How Many) imply an answer of type *Quantity*, etc. The patterns in Esfinge matched with 43.5% of the questions in 2006. For these questions, SIEMÊS tags the relevant text passages, in order to count the number of occurrences of NEs belonging to the relevant categories. SIEMÊS was used in 54% of the answered questions in the PT-PT task this year.

For other questions, however, either the answer is not a named entity (definition questions) or it is very time consuming to create patterns to deal with them (questions of type *Qual X / Which X*, where *X* can be potentially anything). For these questions (or when the estimated number of answers for the question are not obtained within the candidate answers extracted with the NER), Esfinge uses the n-grams module which counts the word n-grams in the relevant text passages. In previous years we used an external tool to harvest n-grams, but since we were using a very limited set of functionalities from that tool, this year we developed our own module in order to make the installation of Esfinge less dependent from external tools.

The candidate answers (obtained either through NER or n-gram harvesting) are then ranked according to their frequency, length and the score of the passage from where they were retrieved using the formula:

Candidate answer score = $\sum (F * S * L)$, through the passages retrieved in the previous modules where:

F = Frequency of candidate
 S = Score of the passage
 L = Length of candidate

This year we managed to install SIEMÊS in our server which enabled us to analyze all the relevant passages with this NER system. In average, 30 different named entities of the relevant types were identified in the passages retrieved from the Web for each question, while 22 different named entities of the relevant types were identified in the passages retrieved from the document collection. Last year, it was not possible to analyze the whole relevant passages for practical reasons, so only the most frequent sequences of 1 to 3 words extracted from these relevant passages were analyzed which meant that the NER system was used in quite sub-optimal conditions. Nevertheless, this sub-optimal use of the NER led to some improvement in the results, as described in [9]. This year we expected still larger improvement.

2.5 Database of Co-occurrences

The score computed in the previous modules is heavily dependent on word frequency though, therefore, very frequent words may appear quite often in the results. The use in a question answering system of an auxiliary corpus to check the frequency of words in order to capture an overrepresentation of candidate answers in a set of relevant passages was successfully tested before; see [10] for instance. With that in mind, we used the database of word co-occurrences BACO [5] this year, to take into account the frequency in which n-grams appear in a large corpus. BACO includes tables with

the frequencies of word n-grams from length 1 to 4 of the Portuguese Web collection WPT-03, a snapshot of the Portuguese Web in 2003 (around 10^9 words) [11]. The scores obtained in the previous module are then adjusted, giving more weight to more rare candidates using the formula:

Candidate answer adjusted score = Candidate answer score * log (Number of words in BACO / Candidate answer frequency in BACO¹)

3 Official Runs

As in previous years we grouped the questions in a set of categories in order to get a better insight on which type of questions the system gets better results. For example, the category *Place* includes the questions where the system expects to have a location as answer such as the *Onde* (Where) questions and the *Que X* (Which X) questions where $X=cidade$ (city) or $X=país$ (country), in which the NER system is used to identify names of countries, cities or other locations in the retrieved documents. Other interesting categories are *People*, *Quantities* and *Dates* where the NER system is also used to find instances of those categories in the relevant texts.

Other categories are more pattern-oriented like the categories *Que é X* (What is X) or *QualQual X* (Which X).

Table 1. Results of Esfinge by type of question in 2005 and 2006

Type of question (after semantic analysis)		No. of Q. in 2006	No. (%) of correct answers			No. of Q. in 2005	No. (%) of correct Answers 2005	
			Run 1 PT-PT	Run 2 PT-PT	Run EN-PT		Best PT-PT Run	Run EN-PT
NER	People	29	9 (31%)	8 (28%)	2 (7%)	47	11 (23%)	5 (11%)
	Place	26	11 (42%)	10 (38%)	8 (31%)	33	9 (27%)	2 (6%)
	Date	20	3 (15%)	3 (15%)	1 (5%)	15	3 (20%)	2 (13%)
	Quantity	7	1 (14%)	1 (14%)	1 (14%)	16	4 (25%)	1 (6%)
	Height, Area, Duration, Distance, Organization	5	1 (20%)	1 (20%)	0	4	1 (25%)	0
n-grams	QualQual X ¹	60	11 (18%)	10 (17%)	9 (15%)	34	8 (24%)	6 (17%)
	Que é X ²	36	10 (28%)	11 (31%)	4 (11%)	15	2 (13%)	0
	Quem é <HUM> ³	9	3 (33%)	2 (22%)	4 (44%)	27	6 (22%)	6 (22%)
	Como se chama / Diga X ⁴	8	1 (13%)	0	0	9	4 (44%)	2 (22%)
Total		200	50 (25%)	46 (23%)	29 (15%)	200	48 (24%)	24 (12%)

¹) Which X, ²) What is X, ³) Who is <HUM>, ⁴) What is X called / Name X

¹ The candidate answer frequency in BACO is set to 1 when there are no occurrences in BACO.

Table 1 summarizes the results of the three official runs. Two runs which used the algorithm described in section 2 were sent in the PT-PT task. The only difference is that Run 1 used word n-grams from length 1 to 3, whereas in Run 2, word n-grams from length 1 to 4 were used. The EN-PT run used word n-grams from length 1 to 3.

This year there was a considerable increase in the number of definition questions of type *Que é X?* and a considerable reduction in the number of questions of type *Quem é <HUM>?*. Theoretically this evolution would make Esfinge's task more complicated, as it was precisely for the questions of type *Que é X?* that the system obtained the worst results last year.

However, Esfinge managed to have even a slightly better result than last year in Run 1. This was accomplished through considerable improvements in the results with the aforementioned questions of type *Que é X?*, as well as with the answers of type *People* and *Place*.

In Run 2, Esfinge managed to answer correctly some questions for which it was not possible to have a complete answer in Run 1 (Run 1 used word n-grams up to length 3, so when the n-gram module was used, the maximum answer length was 3, and, therefore, it was not possible to give an exact answer to questions that required lengthier answers). However, as Run 2 used longer n-grams, 7 of the answers that were returned even though including the correct answer, included also some extra words. As these answers were classified as Inexact, the global results of Run 2 are slightly worse than the ones in Run 1. Not surprisingly, definitions of type *Que é X?* were the only type of questions with some improvement. If we consider the correct answers plus the answers including the correct answer and some more words then we obtain 51 answers in Run 1 and 53 answers in Run 2.

Regarding the EN-PT there was a slight improvement in the number of exact answers, even though only small adjustments were made in the automatic translation of some questions.

Table 2 summarizes the main causes for wrong answers. As in last year, most of the errors in the PT-PT task reside in the document retrieval. More than half of the errors occurred because the system was not able to find relevant documents to answer the questions.

The other two main causes for errors were the answer scoring algorithm and the answer supporting module. The category Others includes among other causes: the answer needing more than four words (only word n-grams up to length four are considered), missing patterns to relate questions to answer categories, NER failures and the retrieved documents from the Web not including any answer.

Table 2. Causes for wrong answers

Problem	No. of wrong answers		
	Run 1 PT-PT	Run 2 PT-PT	Run EN-PT
Translation	-	-	106 (62%)
No documents retrieved in the document collection	84 (56%)	84 (55%)	25 (15%)
Answer scoring algorithm	37 (24%)	30 (19%)	19 (11%)
Answer support	10 (7%)	11 (7%)	13 (8%)
Others	19 (13%)	29 (19%)	8 (5%)
Total	150	154	171

4 Beyond the Official Runs

In addition to the two submitted monolingual runs, some more experiments were performed. Table 3 gives an overview of the results obtained in these experiments. Run 3 used the same algorithm used in Run 1 except that it did not use the BACO database of word co-occurrences. Run 4 did not use the Web, only the document collection was used to extract the answers.

Table 3. Results of non-official runs (Runs 3 and 4) and the best official run (Run 1) of Esfinge

Type of question		No. of Q. in 2006	No. (%) of exact answers		
			Run 3	Run 4	Run 1
NER	People	29	7 (24%)	6 (21%)	9 (31%)
	Place	26	10 (38%)	10 (38%)	11 (42%)
	Date	20	3 (15%)	3 (15%)	3 (15%)
	Quantity	7	1 (14%)	1 (14%)	1 (14%)
	Height, Duration, Area, Organization, Distance	5	1 (20%)	0	1 (20%)
n-grams	Que/Qual X	60	11 (18%)	9 (15%)	11 (18%)
	Que é X	36	10 (28%)	3 (8%)	10 (28%)
	Quem é <HUM>	9	3 (33%)	0	3 (33%)
	Como se chama/ Diga X	8	1 (13%)	1 (13%)	1 (13%)
Total		200	47 (24%)	33 (17%)	50 (25%)

From the results of Run 3, one can conclude that the use of the database of word co-occurrences to adjust the ranking of the candidate answers did not improve the results in a significant way, but it is somehow surprising that the questions correctly answered in the official run that were not answered correctly in Run 3 are precisely questions with answers of type *Person* and *Place*, obtained with the NER recognizer. One would expect that the use of BACO would be more useful with answers obtained using n-gram harvesting.

The results of Run 4 are quite interesting on the other hand. While the best run using the Web achieved an accuracy of 25%, this experiment with exactly the same algorithm except that it does not use the Web achieved only an accuracy of 17%. This is a bit surprising since in last year experiments the difference was not this large (24% using the Web and 22% with the document collection only as reported in [4]). Another issue concerns the fact that the document collection is getting older and therefore it should be more difficult to find a good answer in the Web (which is constantly updated with new information) for some of the questions that one can make based in this collection. Questions like *Quem é o presidente de X?* (Who is the president of

$X?$) or *Quem é <HUM>?* are examples of questions where the Web can give some steadily noisier information as the collection gets older. The fact that more definitions of type *Que é X?* were included this year at the expense of questions of type *Quem é <HUM>?* can explain this somehow surprising result of Run 4.

5 What About Using the Web to Support Answers?

As seen in Table 2, Esfinge is not able to find relevant documents for more than half of the questions. This means that our retrieval patterns are not good enough and we could try to improve them either manually or semi-automatically (pattern evaluation, machine learning, etc.).

However, one of the main purposes of the work with Esfinge is to study whether using the information redundancy existent in the Web enables us to use simpler approaches. This section describes preliminary experiments to evaluate what can be achieved using the Web not only to retrieve answers, but also to support them.

We did an experiment with the 101 questions for which the system was not able to find relevant documents in the CLEF collection. In this experiment, Esfinge was configured to support answers with Web URLs and Web passages (as usual Google and Yahoo were used to retrieve text from the Web), instead of docids and passages from documents in the CLEF collection. Table 4 shows the number of R(ight), U(nsupported), ine(X)act and W(rong) answers obtained for each type of question.

Table 4. Results obtained using the Web to support answers

Type of question (after semantic analysis)		No. of questions	R	U	X	W
NER	People	17	1 (6%)	1 (6%)	1 (6%)	14 (82%)
	Place	13	2.6 (20%)	0	1 (8%)	9.4 (72%)
	Date	12	5 (42%)	0	0	7 (58%)
	Quantity	4	0	0	0	4 (96%)
n-grams	QueQual X	49	1 (2%)	0	1 (2%)	47 (80%)
	Que é X	5	0	1 (20%)	0	4
	Como se chama/ Diga X	1	1 (100%)	0	0	0
Total		101	10.6 (10%)	2 (2%)	3 (3%)	85.4 (85%)

Using the Web, Esfinge is able to answer correctly 10 of the answers for which it was not able to find relevant documents in CLEF collection. Nevertheless, the findings in the error analysis were even more promising: for more than half of the questions wrongly answered, even though not finding the right answer, Esfinge manages to find relevant documents. For 21% of the questions no documents were retrieved at all whereas only non relevant documents were retrieved for 25% of the questions.

This error analysis was also very useful because while doing it we found some errors that were somehow hidden due to our dual strategy: find answers in the Web and then try to project them in CLEF document collection. Focusing on what the system is

doing while retrieving passages from the Web enabled us to find some details that can be improved: text from the Web, even though helping sometimes, arises some challenges not found in the CLEF document collection.

Another interesting issue is the NIL questions (questions that do not have an answer in the target language in CLEF collection). For Portuguese there were 18 of these questions, but Esfinge managed to find relevant documents for 10 of these questions in the Web, gave one (U)nsupported answer to one of these questions and one ine(X)act answer to another one.

6 Concluding Remarks

I think that the QA task at CLEF is moving in the right direction. This year's edition had some novelties that I consider very positive. For example, the fact that this year systems were required to provide document passages to support their answers, whereas in last year the document ID sufficed, makes the task more user-centered and realistic. It is not realistic to assume that a real user would be satisfied with an answer justification that consisted in a (sometimes) quite large document. Another positive innovation was not to provide the type of question to the systems.

The main progress in Esfinge this year consisted in the refactoring of the code produced in previous years that is now available to the community in <http://www.linguatca.pt/Esfinge/>. As to the results of the runs themselves, the system managed to maintain the same level of performance obtained last year, even though questions were more challenging this year (the accuracy of this year's system with 2004 and 2005 questions is 34% and 37% respectively). It was also interesting to notice that the use of the Web as an auxiliary resource proved to be more crucial than in previous years. This might indicate that either more information (and more reliable) is retrieved by Google and Yahoo in Portuguese, or that the questions this year were less time-dependent. The results of the experiment described in section 5 underline also the importance that the Web can have in question answering: the performance of an algorithm can be improved simply by using the Web to a fuller extent.

I think that for CLEF 2007 we could eliminate the NIL questions and open the possibility for Web based support for the answers. The NIL questions being used are not that challenging (most of them are just created around facts that exist in a certain language document collection and that do not exist in other languages document collections). Allowing Web based justifications would enable the creation of other interesting questions, not being constantly trapped in 1994/1995 (the document collections we have been using).

Acknowledgments. I thank Diana Santos for reviewing previous versions of this paper, Luís Sarmiento for providing support for the use of both SIEMÊS and BACO.

The work reported here has been (partially) funded by Fundação para a Ciência e Tecnologia (FCT) through project POSI/PLP/43931/2001, co-financed by POSI, and by POSC project POSC/339/1.3/C/NAC.

References

1. Brill, E.: Processing Natural Language without Natural Language Processing. In: Gelbukh, A. (ed.) *CICLing 2003*. LNCS, vol. 2588, pp. 360–369. Springer, Heidelberg (2003)
2. Gomes, D., Silva, M.J.: Characterizing a National Community Web. *ACM Transactions on Internet Technology (TOIT)* 5(3), 508–531 (2005)
3. Costa, L.: First Evaluation of Esfinge - a Question Answering System for Portuguese. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004*. LNCS, vol. 3491, pp. 522–533. Springer, Heidelberg (2005)
4. Costa, L.: 20th Century Esfinge (Sphinx) solving the riddles at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 467–476. Springer, Heidelberg (2006)
5. Sarmiento, L.: BACO - A large database of text and co-occurrences. In: *Proceedings of LREC*, May 22–28, Genoa, Italy (2006)
6. Sarmiento, L.: SIEMÊS - a named entity recognizer for Portuguese relying on similarity rules. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) *PROPOR 2006*. LNCS (LNAI), vol. 3960, pp. 90–99. Springer, Heidelberg (2006)
7. Bick, E.: The Parsing System “Palavras”: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, Aarhus (2000)
8. Mota, C., Santos, D., Ranchhod, E.: Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. In: Santos, D. (ed.) *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, pp. 161–175. IST Press (2007)
9. Costa, L., Sarmiento, L.: Component Evaluation in a Question Answering System. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., Tapias, D. (eds.) *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC’2006)*, Genoa, Italy, 22–28 May 2006, pp. 1520–1523 (2006)
10. Clarke, C.L.A., Cormack, G.V., Lynam, T.R.: Exploiting Redundancy in Question Answering. In: *Research and Development in Information Retrieval*, pp. 358–365 (2001)
11. Cardoso, N., Martins, B., Gomes, D., Silva, M.J.: WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa. In: Santos, D. (ed.) *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, pp. 279–288. IST Press (2007)

Using Machine Learning and Text Mining in Question Answering

Antonio Juárez-González, Alberto Téllez-Valero, Claudia Denicia-Carral,
Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
{antjug, albertotellezv, cdenicia, mmontesg, villasen}@inaoe.mx

Abstract. This paper describes a QA system centered in a full data-driven architecture. It applies machine learning and text mining techniques to identify the most probable answers to factoid and definition questions respectively. Its major quality is that it mainly relies on the use of lexical information and avoids applying any complex language processing resources such as named entity classifiers, parsers and ontologies. Experimental results on the Spanish Question Answering task at CLEF 2006 show that the proposed architecture can be a practical solution for monolingual question answering by reaching a precision as high as 51%.

1 Introduction

Current information requirements suggest the need for efficient mechanisms capable of interacting with users in a more natural way. Question Answering (QA) systems have been proposed as a feasible option for the creation of such mechanisms [1]. Recent developments in QA use a variety of linguistic resources to help in understanding the questions and the documents. The most common linguistic resources are: part-of-speech taggers, parsers, named entity extractors, dictionaries, and WordNet [2, 3, 4, 5, 10]. Despite the promising results of these approaches, they have two main inconveniences. On the one hand, the construction of such linguistic resources is a very complex task. On the other hand, their performance rates are usually not optimal.

In this paper we present a QA system that can answer factoid and definition questions. This system is based on a full data-driven approach that requires a minimum knowledge about the lexicon and the syntax of the specified language. It is built on the idea that the questions and their answers are commonly expressed using the same set of words. Therefore, it simply uses lexical information to identify the relevant document passages and to extract the candidate answers.

The proposed system continues our previous work [6] by considering a lexical data-driven approach. However, it presents two important modifications. First, it applies a supervised approach instead of a statistical method for answering factoid questions. Second, it answers definition questions by applying lexical patterns that were automatically constructed rather manually defined.

The following sections give some details of the system. In particular, section 2 describes the method for answering factoid questions, section 3 explains the method for answering definition questions, and section 4 discusses the results achieved by our system at the CLEF 2006 Spanish Question Answering task.

2 Answering Factoid Questions

Figure 1 shows the general process for answering factoid questions. It considers three main modules: *passage retrieval*, where the passages with a high probability of containing the answer are recovered from the document collection; *question classification*, where the type of expected answer is determined; and *answer extraction*, where candidate answers are selected using a machine-learning approach, and the final answer recommendation of the system is produced. The following sections describe each of these modules.

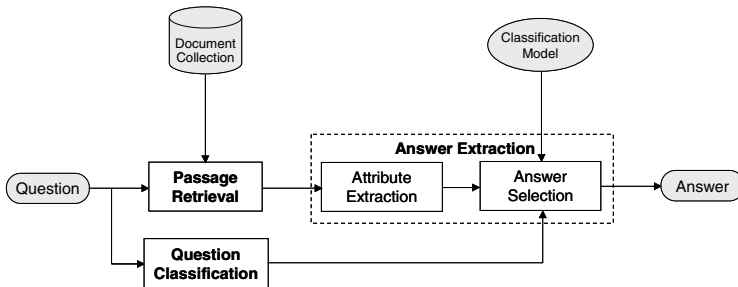


Fig. 1. Process for answering factoid questions

2.1 Passage Retrieval

The passage retrieval (PR) method is especially suited to the QA task. It allows passages with the highest probability of containing the answer to be retrieved, instead of simply recovering the passages sharing a subset of words with the question.

Given a user question, the PR method finds the passages with the relevant terms (non-stopwords) using a classical information retrieval technique based on the vector space model. Then, it measures the similarity between the n -gram sets of the passages and the user question in order to obtain the new weights for the passages. The weight of a passage is related to the largest n -gram structure of the question that can be found in the passage itself. The larger the n -gram structure, the greater the weight of the passage. Finally, it returns to the user the passages with highest weights.

Details about the PR method can be found in [7].

2.2 Question Classification

Given a question, this module is responsible for determining the semantic class of the expected answer. This information will be used later to reduce the searching space.

The idea is to focus the answer extraction only on those text fragments related to the expected type of answer.

Our system prototype implements this module following a direct approach based on regular expressions. It only considers three general semantic classes of expected answers: dates, quantities and names (i.e., general proper nouns).

2.3 Answer Extraction

Answer extraction aims to establish the best answer for a given question. It is based on a supervised machine-learning approach. It consists of two main modules, one for attribute extraction and the other for answer selection.

Attribute extraction. First, the recovered passages are processed in order to identify all text fragments related to the expected type of answer. Each one of the identified text fragments is considered as a “candidate answer”. It is important to mention that this analysis is also based on a set of regular expressions.

In a second step, the lexical context of each candidate answer is analyzed with the aim of constructing its formal representation. In particular, each candidate answer is represented by a set of 17 attributes, clustered in the following groups:

1. Attributes that describe the complexity of the question, for instance, its length (number of non-stopwords).
2. Attributes that measure the similarity between the context of the candidate answer and the given question. Some of them describe the word overlap between the question and the context of the candidate answer. Some others measure the density of the question words in the context of the candidate answer.
3. Attributes that indicate the relevance of the candidate answer in reference to the set of recovered passages. Some of these attributes are: the redundancy of the candidate answer in the set of recovered passages, and the position of the passage containing the candidate answer.

Answer Selection. This module is based on a machine-learning approach. Its purpose is to select, from the set of candidate answers, the one with the maximum probability of being the correct answer. In particular, this module is implemented by a Naïve Bayes classifier, which was constructed using as training set the questions and documents from the previous CLEF campaigns.

3 Answering Definition Questions

Figure 2 shows the general scheme of our method for answering definition questions. It consists of three main modules: a module for the discovery of definition patterns, a module for the construction of a general definition catalogue, and a module for the extraction of the candidate answer. The following sections describe in detail these modules.

It is important to mention that this method is especially suited to answering definition questions as defined in CLEF. That is, questions asking for the position of a person,

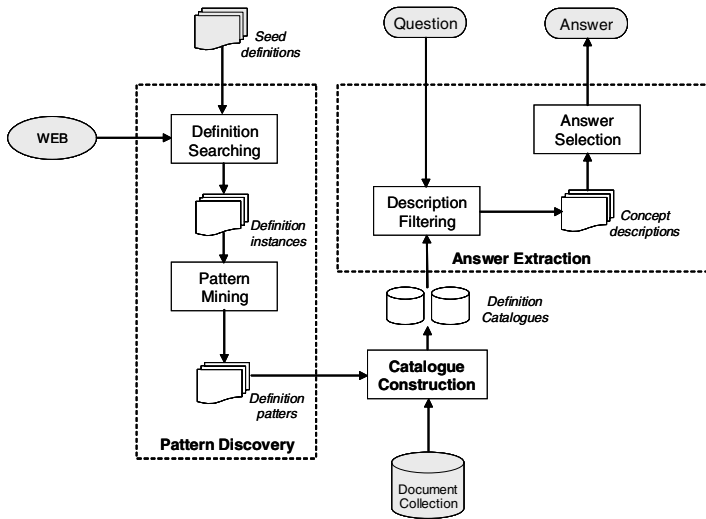


Fig. 2. Process for answering definition questions

e.g., Who is Vicente Fox?, and for the description of concept, e.g., What is CERN? or What is Linux?.

It is also important to notice that the processes for pattern discovery and catalogue construction are done offline, while the answer extraction is done online, and that in contrast to traditional QA approaches, the proposed method does not use any module for document or passage retrieval.

3.1 Pattern Discovery

The module for pattern discovery uses a small set of concept-description pairs to collect from the Web an extended set of definition instances. Then, it applies a text mining method to the collected instances to discover a set of definition surface patterns. The idea is to capture the definition conventions through their repetition. This module involves two main subtasks:

Definition searching. This task is triggered by a small set of empirically defined concept-description pairs. The pairs are used to retrieve a number of usage examples from the Web¹. Each usage example represents a definition instance. To be relevant, a definition instance must contain the concept and its description in one single phrase.

Pattern mining. It is divided into three main steps: data preparation, data mining and pattern filtering. The purpose of the data preparation phase is to normalize the input data. It transforms all definition instances into the same format using special tags for the concepts and their descriptions. It also indicates with a special tag the concepts expressing proper names.

¹ At present we are using Google for searching the Web.

In the data mining phase, a sequence mining algorithm [9] is used to obtain all maximal frequent sequences of words², punctuation marks and tags from the set of definition instances. The sequences express lexicographic patterns highly related to concept definitions.

Finally, the pattern-filtering phase allows choosing the more discriminative patterns. It selects the patterns satisfying the following general regular expressions:

```

<left-string> DESCRIPTION <middle-string> CONCEPT <right-string>
<left-string> CONCEPT <middle-string> DESCRIPTION <right-string>
<left-string> DESCRIPTION <middle-string> PROPER_NAME_CONCEPT <right-string>
<left-string> PROPER_NAME_CONCEPT <middle-string> DESCRIPTION <right-string>
<left-string> DESCRIPTION <middle-string> PROPER_NAME_CONCEPT
PROPER_NAME_CONCEPT <middle-string> DESCRIPTION <right-string>
<left-string> DESCRIPTION PROPER_NAME_CONCEPT
PROPER_NAME_CONCEPT DESCRIPTION <right-string>

```

The idea of the pattern discovery process is to obtain several surface definition patterns, starting with a small set of concept-description example pairs (details about this process can be found in [8]). For instance, using description seeds “*Wolfgang Clement – German Federal Minister of Economics and Labor*” and “*Vicente Fox – President of Mexico*”, at the end we could obtain definition patterns such as “, the <DESCRIPTION>, <CONCEPT>, says” and “the <DESCRIPTION> <PROPER_NAME_CONCEPT>”. It is important to notice that the discovered patterns may include words and punctuation marks as well as proper name tags as frontier elements.

3.2 Catalogue Construction

In this module, the discovered definition patterns are applied over the target document collection. The result is a set of matched text segments that presumably contain a concept and its description. The definition catalogue is created gathering all matched segments.

3.3 Answer Extraction

This module handles the extraction of the answer for a given definition question. Its purpose is to find the most adequate description for a requested concept from the definition catalogue. The definition catalogue may contain a huge diversity of information, including incomplete and incorrect descriptions for many concepts. However, it is expected the correct information will be more abundant than incorrect information. This expectation supports the idea of using a frequency criterion and a text mining technique to distinguish between the adequate and the improbable answers to a given question. This module considers the following steps:

Description filtering. Given a specific question, this procedure extracts from the definition catalogue all descriptions corresponding to the requested concept. As we mentioned, these “presumable” descriptions may include incomplete and incorrect

² The word sequence p is frequent in the collection D if it occurs in at least σ texts of D , and it is maximal if there does not exist any sequence p' in D such that p is a subsequence of p' and p' is also frequent in D .

information. However, it is expected that many of them will contain, maybe as a substring, the required answer.

Answer selection. This process aims to detect a single answer to the given question from the set of extracted descriptions. It is divided into two main phases: data preparation and data mining.

The data preparation phase focuses on homogenizing the descriptions related to the requested concept. The main action is to convert these descriptions to a lower case format. In the data mining phase, a sequence mining algorithm [9] is used to obtain all maximal frequent word sequences from the set of descriptions. Finally, each sequence is ranked based on the frequency of occurrence of its subsequences in the whole description set [8]. The best one is given as the final answer.

Figure 3 shows the process of answer extraction for the question “Who is Diego Armando Maradona?”. First, we obtained all descriptions associated with the requested concept. It is clear that there are erroneous or incomplete descriptions (e.g. “Argentina soccer team”). However, most of them contain a partially satisfactory explanation of the concept. Actually, we detected correct descriptions such as “captain of the Argentine soccer team” and “Argentine star”. Then, a mining process allowed detecting a set of maximal frequent sequences. Each sequence was considered a candidate answer. In this case, we detected three sequences: “argentine”, “captain of the Argentine soccer team” and “supposed overuse of Ephedrine by the star of the Argentine team”. Finally, each candidate answer was ranked based on the frequency of occurrence of its subsequences in the whole description set. In this way, we took advantage of the incomplete descriptions of the concept. The selected answer was “captain of the Argentine national football soccer team”, since it was formed from frequent subsequences such as “captain of the”, “soccer team” and “argentine”.

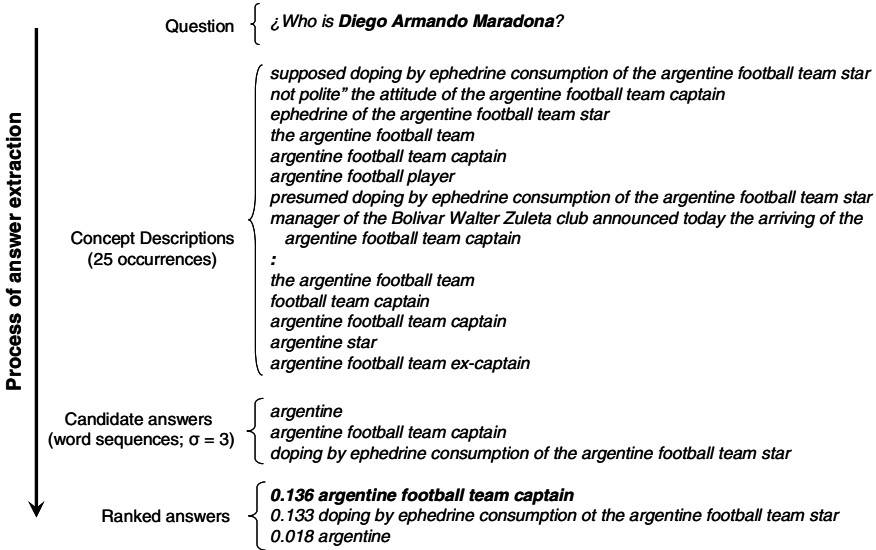


Fig. 3. Data flow in the answer extraction process

4 Evaluation Results

This section describes the experimental results related to our participation at the QA@CLEF2006 monolingual track for Spanish. It is important to remember that this year the question type (e.g., factoid, definition, temporal or list) was not included as a data field in the question test file. Therefore, each participant had to automatically determine the kind of question.

Our system prototype, as described in the previous sections, can only deal with factoid and definition questions. From the 200 test questions, it treats 144 as factoid questions and 56 as definition questions. Table 1 details our results on answering both types of questions considering the evaluation given by CLEF judges and our own evaluation.

Table 1. Evaluation of the system on answering factoid and definition questions

	Accuracy	R	W	X	U	Factoid Questions (144)	Definition Questions (56)
CLEF evaluation	51%	102	86	3	9	59 (40.9%)	43 (76.7%)
Local evaluation	53%	106	86	3	5	63 (43.75%)	43 (76.7%)

CLEF evaluation gave our system an accuracy of 51%, by answering 102 questions correctly, 59 factoids and 43 definitions. In contrast, in our local evaluation we obtain an accuracy of 53%. This difference was caused by a discrepancy in the evaluation of four factoid questions about dates. The answers for these questions lacked a year string in the passages, therefore we decided to extract the year string from the document ID. This way, we obtained an entire date answer (consisting of a day, month and a year) completely supported considering both the passage and the document ID. Nevertheless, CLEF judges considered these four questions as unsupported, affecting the overall accuracy of our system. As an example, table 2 shows our answer for one of these questions.

Table 2. An adapted answer for a date question

Document ID	Question	Answer	Passage
EFE19950608-05247	When was the G7 meeting in Halifax?	June 15th to 17th 1995	...the meeting of the seven most industrialized countries (G7) in Halifax (Canada), from June 15th to 17th , and the European Council of Cannes- figure in the order of the day of this informal summit.

Lastly, it is important to point out some facts about our participation at the QA@CLEF 2006 [11]. First, our prototype was one of the four systems that went beyond the barrier of 50% in accuracy. Second, it was the best system for answering

definition questions. Third, the overall evaluation accuracy of this year exercise (51%) was 10-points over our result for last year [8].

5 Conclusions and Future Work

This paper presented a question answering system that can answer factoid and definition questions. This system is based on a lexical data-driven approach. Its main idea is that the questions and their answers are commonly expressed using almost the same set of words, and therefore, it simply uses lexical information to identify the relevant passages as well as the candidate answers.

The answer extraction for factoid questions is based on a machine learning method. Each candidate answer (an uppercase word denoting a proper name, a date or a quantity) is represented by a set of lexical attributes and a classifier determines the most probable answer for the given question. The method achieved good results, however it has two significant disadvantages: (i) it requires a lot of training data, and (ii) the detection of the candidate answers is not always (not for all cases, nor for all languages) an easy task to perform with high precision.

On the other hand, the answer extraction for definition questions is based on a text mining approach. The proposed method uses a text mining technique (namely, a sequence mining algorithm) to discover a set of definition patterns from the Web as well as to determine with finer precision the answer to a given question. The achieved results were especially good, and they showed that a non-standard QA approach, which does not contemplate an IR phase, can be a good scheme for answering definition questions.

As future work we plan to consider syntactic attributes in the process of answering factoid questions. These attributes will capture the matching of syntactic trees as well as the presence of synonyms or other kinds of relations among words. In addition, we are also planning to improve the final answer selection by applying an answer validation method.

Acknowledgments. This work was done under partial support of CONACYT (Project Grant: 43990). We would also like to thank the CLEF organizing committee as well as to the EFE agency for the resources provided.

References

1. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question Answering Track. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, Springer, Heidelberg (2005)
2. Ferrández, S., López-Moreno, P., Roger, S., Ferrández, A., Peral, J., Alvarado, X., Noguera, E., Llopis, F.: AliQAn and BRILI QA Systems at CLEF 2006. In: Working notes for the Cross Language Evaluation Forum Workshop, September 2006, (CLEF 2006) (2006)

3. Buscaldi, D., Gomez, J.M., Rosso, P., Sanchis, E.: The UPV at QA@CLEF 2006. In: Working notes for the Cross Language Evaluation Forum Workshop, September 2006, (CLEF 2006) (2006)
4. de-Pablo-Sánchez, C., González-Ledesma, A., Moreno, A., Martínez-Fernández, J.L., Martínez, P.: MIRACLE at the Spanish CLEF@QA 2006 Track. In: Working notes for the Cross Language Evaluation Forum Workshop, September 2006, (CLEF 2006) (2006)
5. Ferrés, D., Kanaan, S., González, E., Ageno Al, R.H., Turmo, J.: The TALP-QA System for Spanish at CLEF-2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
6. Montes-y-Gómez, M., Villaseñor-Pineda, L., Pérez-Coutiño, M., Gómez-Soriano, J.M., Sanchis-Arnal, E., Rosso, P.: INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
7. Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Villaseñor-Pineda, L., Rosso, P.: Language Independent Passage Retrieval for Question Answering. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, Springer, Heidelberg (2005)
8. Denicia-Carral, C., Montes-y-Gómez, M., Villaseñor-Pineda, L., García-Hernández, R.: A Text Mining Approach for Definition Question Answering. In: Proceedings for the Fifth International Conference on Natural Language Processing (FinTal 2006), Turku, Finland (August 2006)
9. García-Hernández, R., Martínez-Trinidad, F., Carrasco-Ochoa, A.: A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, Springer, Heidelberg (2006)
10. Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, P., Vidal, D.: Prib-eram's question answering system in a cross-language environment. In: Working notes for the Cross Language Evaluation Forum Workshop, September 2006, (CLEF 2006) (2006)
11. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Working notes for the Cross Language Evaluation Forum Workshop, September 2006, (CLEF 2006) (2006)

Applying Dependency Trees and Term Density for Answer Selection Reinforcement

Manuel Pérez-Coutiño¹, Manuel Montes-y-Gómez², Aurelio López-López²,
Luis Villaseñor-Pineda², and Aarón Pancardo-Rodríguez¹

¹ Vanguard Engineering Puebla (VEng)

Recta a Cholula 308-4, CP 72810, San Andrés Cholula, Pue., México

{mapco, apancardo}@v-eng.com

² Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Luis Enrique Erro No. 1, CP 72840, Sta. Ma. Tonantzintla, Pue., México

{mmontesg, allopez, villasen}@inaoep.mx

Abstract. This paper describes the experiments performed for the QA@CLEF-2006 within the joint participation of the eLing Division at VEng and the Language Technologies Laboratory at INAOE. The aim of these experiments was to observe and quantify the improvements in the final step of the Question Answering prototype when some syntactic features were included into the decision process. In order to reach this goal, a shallow approach to answer ranking based on the term density measure has been integrated into the weighting schema. This approach has shown an interesting improvement against the same prototype without this module. The paper discusses the results achieved, the conclusions and further directions within this research.

1 Introduction

Over the last years, Question Answering (QA) system research has shown an incremental growing both in interest as well as in complexity. Particularly, QA for Spanish has been formally evaluated within the CLEF initiative during the last three years, increasing the number of participant groups and, consequently, the proposed systems. Those QA systems have explored different approaches to cope with the QA problem, analyzing several methodologies from purely data-driven [7, 11] to in-depth natural language processing (NLP) [5]. Despite the methods used in these approaches, the improvement of factoid question resolution has been measured.

This paper presents the prototype developed as a shared effort between the recently formed eLing Division at VEng and the Language Technologies laboratory at INAOE. This approach continues with the previous work of the authors [9] to cope with factoid questions resolution. The aim of these experiments was to observe and quantify the possible improvement at the answer selection step of a QA prototype, as a consequence of introducing some syntactic features to the decision process. In order to reach this goal, the following key points have been included in the QA prototype: i) a syntactic parser based on dependency grammars; ii) a shallow technique to weigh the number of question terms which have a syntactic dependency to one candidate answer within a relevant passage to the given question (*term density*); iii) a formula

for merging the term density of each candidate answer and the weights gathered in the previous steps to obtain its total weight. Finally, the answer selection process arranges candidate answers based on their weights, selecting the top- n as the QA system answers. The approach described in this document has shown an interesting improvement up to 15% against the same QA prototype without this module.

The rest of the paper is organized as follows: Section 2 summarizes the architecture of the prototype; section 3 exposes the details of the new elements within the prototype architecture; section 4 discusses the results achieved by the system. Finally, section 5 draws conclusions and discusses further work.

2 Prototype Architecture

As stated before, the system developed for QA@CLEF 2006 is based on the previous work of the authors [9], where the most important modification relies on the inclusion of syntactic features to the decision process at the answer selection module. Figure 1 shows the main blocks of the system. It could be noticed that factoid and definition questions are handled independently. This report is focused on the factoid questions resolution process, while details of the processes involved in the creation and use of definition patterns aimed to answering definition questions can be found in [4].

Factoid question treatment relies on a typical QA architecture, including the following steps: *question processing*, *documents processing*, *searching*, and finally *answer extraction and selection*. Next paragraphs summarize the initial steps of the prototype whilst section three and four discuss the new ones.

2.1 Question Processing

Our prototype implements this step following a straight forward approach, parsing the question with a set of heuristic rules in order to get its semantic class. The MACO POS tagger [3] is used to tag the question, as well as to identify and classify its named entities. This information will be used later, during the searching step, mainly to match questions and candidate answer context, contributing to the weight schema.

2.2 Documents Processing

The processing of target documents first applies MACO [3] to the whole document collection (similar to the question processing). In the second part of the process, and in parallel to the first one, the whole document collection is tagged with the FDG Parser from Conexor, which is based on the *Functional Dependency Grammar* discussed in [12]. The final part of this step is performed by the JIRS [6] passage retrieval system (PRS), which create the index for the searching process. The index gathered by JIRS and the tagged collection are aligned phrase by phrase for each document in the collection. This way, the system could retrieve later the relevant passages for a given question with JIRS, and then use their tagged form for the answer extraction process.

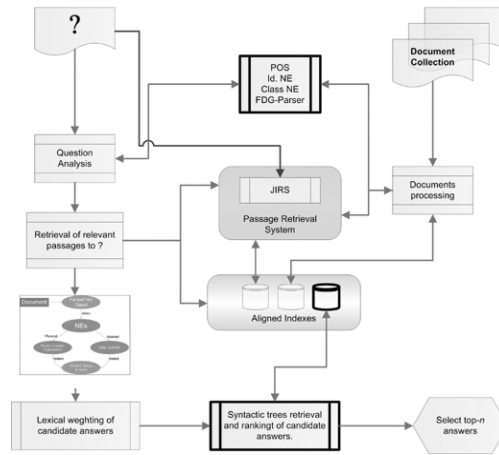


Fig. 1. Block diagram of the system. Factoid and definition questions are treated independently.

2.3 Searching

This step is performed in two parts. The first one consists in retrieving the relevant passages for the given question. This step is performed by JIRS, taking as input the question without previous processing. JIRS ranks the retrieved passages based on the computation of a weight for each passage. This weight is related to the larger n-gram structure of the question that can be found in the passage itself. The larger the n-gram structure, the greater the weight of the passage. Details of JIRS metrics can be found in [6]. The second part of the process requires the POS and Parsing tagged forms of each passage in order to gather the representation used to extract candidate answers. Tagged passages are represented as described in [8] where each retrieved passage is modeled by the system as a factual text object whose content refers to several named entities even when it is focused on a central topic. The model assumes that the named entities are strongly related to their lexical context, especially to nouns (subjects) and verbs (actions). Thus, a passage can be seen as a set of entities and their lexical context. Such representation is used later in order to match question's representation with the best set of candidates gathered from passages.

2.4 Answer Extraction

One of the drawbacks found in our previous work was the lost of precision during answer extraction. Once the system gathers a set of candidate answers, it computes for each candidate answer the lexical weight [9] (see formula 1) in order to rank the best one and then selects the top- n as the system answers. However, there were situations where several candidate answers could have the same weight. In order to minimize such situations syntactic information has been included, computing an additional weight which is later combined with the lexical weight to get the final rank of the candidate answers. This concept is explained in detail in the next section.

$$\omega_{lex}(A) = \frac{t_q}{n} * \left(\frac{|NE_q \cap NE_A|}{|NE_q|} + \frac{|C_q \cap C_A|}{|C_q|} + \frac{F_A(P_i)}{F_A(P)} + \left(1 - \frac{P_i}{k-1} \right) \right) \quad (1)$$

$i=1..k$; k =number of passages retrieved by JIRS

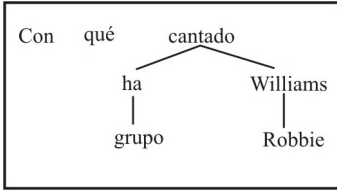
3 Adding Syntactic Features

In order to improve the precision at the answer selection step, we have experimented with the inclusion of a shallow approach based on the use of some syntactic features. The main characteristic of this approach relies on the kind of information used to compute an additional weight to rearrange the set of candidate answers previously selected by a lexical supported method. This approach is flexible given that one of the factors taken in mind was to realize the amount of errors that a parser could yield.

3.1 Key Observations of the Method

There are different approaches to the use of syntactic information within the QA task. Some of them try to found a direct similarity between question structure and those of the candidate answers. This kind of similarity is commonly supported by a set of syntactic patterns which are expected to be matched by the answers. Some systems applying this approach are described in [1, 2, 10]. Other proposals, like the one presented by Tanev et al. [11] apply transformations to the syntactic tree of the question in order to approximate it to the trees of the relevant passages, where it is supposed that an answer is present.

Despite the degree of effectiveness of those approaches, their principal drawback comes up when the parsing is deficient. With this limitation in mind, and trying to



Notice the differences between question and relevant passage structures. Besides, the analysis of the question shows the second case, whilst the passage tree could be seen as a forest. Finally, the tree where the answer is found "*Take That*" does not contain any question terms.

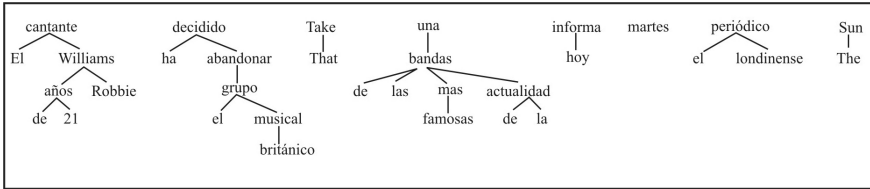
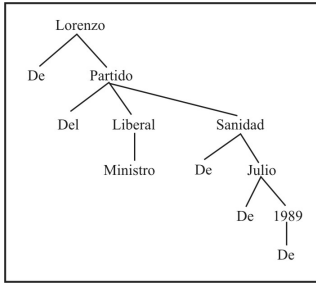


Fig. 2a. Example of syntactic trees gathered from a question and one of its relevant passages



This example shows the tree gathered from the relevant passage to the question: *¿Qué político liberal fue ministro de Sanidad italiano entre 1989 y 1993?*

Notice that the accurate answer “*De Lorenzo*”, occurs within this tree, and relates several question terms to it.

Fig. 2b. Example of one tree with several question terms related to its candidate answer (De Lorenzo)

work around the gap of parsers for Spanish, this proposal is supported by the following observations of the analysis obtained through DFG parser (see figure 2).

1. There are important structural differences between dependency trees of a given question and those gathered from their relevant passages.
2. For a given question, its dependency tree states both, functional and structural dependencies starting from the main verb within the sentence, clearly delimiting other relations like subject, agent, object, etc. Contrasting this fact, dependency trees gathered from the relevant passage of the given question, could be seen as a forest, where functional and structural relations –which could possible lead the process to the accurate answer– are broken from one tree to another.
3. Finally, dependency trees gathered from the relevant passages to a given question could enclose a high number of question terms related to several candidate answers.

3.2 Term Density

In order to cope with these observations, we propose a straight forward metric to capture and weigh up the question terms which are nearest to a candidate answer within a relevant passage. Formula 2 shows those relations, which we have named the *term density* within a dependency tree.

The algorithm applied in order to compute the term density for a candidate answer involves the following steps.

1. For each relevant passage of a given question
 - 1.1. Retrieve the dependency tree for that passage
 - 1.2. For each candidate answer within the passage
 - 1.2.1. Retrieve the sub tree where the candidate answer occurs
 - 1.2.2. Apply formula 2 (δ_q)
 - 1.2.3. Compute the maximum δ_q for each candidate answer, then preserve it if it is greater than 0.5, in other case, $\delta_q=0$

Given:

$Q = \{t_1, t_2 \dots t_n\}$, the questions' terms (lemmas)

$W = \{w_1, w_2 \dots w_k\}$, the passages' terms (lemmas)

$C_i \subset W$, the terms (lemmas) of the i -th candidate answer.

A sub tree $x_\alpha(w_1, \dots, *, \dots, w_h)$: $w_1 \dots w_h$ depends on x_α

(2)

$$\text{Then, } \delta_q(C_i) = \frac{1}{n} \sum_{t_j \in Q} f(t_j) \Leftrightarrow \forall w \in C_i, \langle w, x_\alpha \rangle;$$

$$f(t_j) = \begin{cases} 1, \langle t_j, x_\alpha \rangle \delta t_j(*) \\ 0, \text{else} \end{cases}$$

Finally, computes the total weight of each candidate answer by formula 3.

$$\omega(C_i) = \alpha \omega_{lex}(C_i) + \beta \delta_q(C_i) \quad (3)$$

Where ω_{ex} is the result of computing formula 1; α and β coefficients has been selected experimentally, giving them values $\alpha = 1/3$, and $\beta = 2/3$. This way the syntactic weight has a greater confidence.

4 Experiments and Results

Through experiments performed over several training sets, including the evaluation set of last year QA@CLEF, it could be observed a significant improvement in the final answer selection step. For the case of training with the QA@CLEF-2005 evaluation set, the accuracy of the system was increased over 7% for factoid questions, whilst the improvement in temporal restricted factoid questions achieved a 15%. These percentages represent a good progress giving that over the last years the rate of improvements in the results of Spanish factoid questions evaluation has been gradually. Table 1 shows some examples of the increasing rank for candidate answers with high term density.

Table 1. Examples of rank increasing for candidates' answers with high term density

Question No.	Candidate and Right Answer	Previous Rank TO New Rank	$w_{lex}(C_i)$	$\delta_q(C_i)$	$w(C_i)$
29	De Lorenzo	5 th TO 1 st	0.5472	0.5000	0.5157
115	64 (días)	15 th TO 1 st	0.5440	0.5714	0.5622
139	Yoweri Kaguta Museveni	10 th TO 1 st	0.9367	0.8888	0.9048
161	Jacques Delors	3 rd TO 1 st	0.8505	0.8000	0.8168

The evaluation of QA systems at CLEF 2006 campaign included factoid, definition, temporal and list questions. The organizers provide to participants with a set of 200 unclassified questions, i.e. there were not markers to indicate the type of expected

answer. Another novelty was that teams must provide answers with the specific passage where the answer was extracted from. The later was used to facilitate the evaluation of answers.

We participate in the evaluation with one run. The configuration applied considers three classes of possible answers, Date, Quantity and Proper Nouns; the system analyzes the first 100 1-line passages retrieved by JIRS; the lexical context used was formed with nouns, named entities, verbs, and adjectives, and the size of the window context is 8 words. Table 2 shows the results of the evaluation.

Despite the fact that our results (for factual questions) were only over 2% better than last year, we believe that the approach described could be a good starting point to the introduction of syntactic information to the answer selection process. Some errors observed while training include the confusion of accurate answers by candidate answers that have a term density similar or greater to the right answer. Next direction in our research could include the use of structural relations to get a better discrimination of candidate answers.

Table 2. Results of evaluation at QA@CLEF2006

Run	<i>Vein061eses</i>
Right	80 (45F + 35D + 0 TRF)
Wrong	102
ineXact	3
Unsupported	5
Overall Accuracy	42.11%
Factoid Questions	30.82%
Definition Questions	83.33%
Temporally Restricted Factoid Questions	0%
Overall Confidence Weighted Score (CWS) over F, D, and TR	0.33582

5 Conclusions

This work has presented a method for the inclusion of syntactic information within the final process of answer selection in a QA system. The approach relies in the use of a flexible metric which allows measuring the amount of question terms which have a syntactic dependence to the candidate answers. Although official evaluation does not reach the expectation, preliminary results have demonstrated a significant improvement in the answer selection step. This leads us to think that it could be possible to apply syntactic information in several ways in order to cope with the problem of partial or even more, deficient syntactic trees (in particular dependence trees).

Despite the low increasing in the official evaluation, the methods applied at different steps of the QA process are stable; this conclusion can be inferred from the fact that the prototype has reached its last year performance.

It is important to realize that the additions to our QA prototype presented in this document are limited by the previous processes. This means that the proposed method is not able to extract new candidate answers. Therefore our next steps into QA systems development must be done in the direction of improving system's recall.

Acknowledgements. This work was done under partial support of Vanguard Engineering Puebla SA de CV, CONACYT (Project Grants U39957-Y and 43990), SNI-Mexico, and the Human Language Technologies Laboratory at INAOE. We also like to thanks to the CLEF as well as EFE agency for the resources provided.

References

1. Aunimo, L., Kuuskoski, R.: Question Answering Using Semantic Annotation. In: Peters, C. (ed.) 6th Workshop of the Cross-Language Evaluation Forum, Vienna, Austria (2005)
2. Bertagna, F., Chiran, L., Simi, M.: QA at ILC-UniPi: Description of the Prototype. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, Springer, Heidelberg (2005)
3. Carreras, X., Padró, L.: A Flexible Distributed Architecture for Natural Language Analyzers. In: Proceedings for the LREC'02, Las Palmas de Gran Canaria, Spain (2002)
4. Denicia-Carral, C., Montes-y-Gómez, M., Villaseñor-Pineda, L., García-Hernández, R.: A Text Mining Approach for Definition Question Answering. In: Proceedings for the Fifth International Conference on Natural Language Processing, Turku, Finland (2006)
5. Ferrés, D., Kanaan, S., González, E., Ageno, A., Rodríguez, H.: The TALP-QA system for Spanish at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
6. Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Rosso, P.: A Passage Retrieval System for Multilingual Question Answering. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, Springer, Heidelberg (2005)
7. Montes-y-Gómez, M., Villaseñor-Pineda, L., Pérez-Coutiño, M., Gómez-Soriano, J.M., Sanchis-Arnal, E., Rosso, P.: INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. Pérez-Coutiño, M., Solorio, T., Montes-y-Gómez, M., López-López, A., Villaseñor-Pineda, L.: Toward a Document Model for Question Answering Systems. In: Favela, J., Menasalvas, E., Chávez, E. (eds.) AWIC 2004. LNCS (LNAI), vol. 3034, pp. 145–154. Springer, Heidelberg (2004)
9. Pérez-Coutiño, M., Montes-y-Gómez, M., López-López, A., Villaseñor-Pineda, L.: Experiments for tuning the values of lexical features in Question Answering for Spanish. In: Peters, C. (ed.) 6th Workshop of the Cross-Language Evaluation Forum, Vienna, Austria (2005)
10. Roger, S., Ferrández, S., Ferrández, A., Peral, J., Llopis, F., Aguilar, A., Tomás, D.: AliQAn, Spanish QA System at CLEF-2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
11. Tanev, H., Kouylekov, M., Magnini, B., Negri, M., Simov, K.: Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005. In: Peters, C. (ed.) 6th Workshop of the Cross-Language Evaluation Forum, Vienna, Austria (2005)
12. Tapanainen, P., Järvinen, T.: A Non-Projective Dependency Parser. In: Proceedings for the 5th Conference on Applied Natural Language Processing (1997)
13. Vicedo, J.L., Izquierdo, R., Llopis, F., Muñoz, R.: Question Answering in Spanish. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, Springer, Heidelberg (2004)

Interpretation and Normalization of Temporal Expressions for Question Answering

Sven Hartrumpf and Johannes Leveling

Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen), 58084 Hagen, Germany
`firstname.lastname@fernuni-hagen.de`

Abstract. The German question answering (QA) system InSicht participated in QA@CLEF for the third time. InSicht implements a deep QA approach: it builds on full sentence parses, inferences on semantic representations, and matching semantic representations derived from questions and documents. InSicht was improved for QA@CLEF 2006 as follows: temporal expressions are normalized and temporal deictic expressions are resolved to explicit date representations; the coreference module was extended by a fallback strategy for increased robustness; equivalence rules can introduce negated relations; answer candidates are clustered in order to avoid multiple occurrences of one real-world entity in the answers to a list question; and finally a shallow QA subsystem that produces a second answer stream was integrated into InSicht. The current system is evaluated in an ablation study on the German questions from QA@CLEF 2006.

1 Introduction

The German question answering (QA) system InSicht participated in QA@CLEF for the third time. InSicht implements a deep (or: semantic) approach because it builds on full sentence parses, inferences on semantic representations, and matching between semantic representations derived from questions and document sentences. For QA@CLEF 2005, a coreference resolution module was integrated and several large knowledge bases containing relations between lexical concepts were added [1]. These knowledge bases were generated mainly from nominal compound analyses. InSicht was improved in the following directions for QA@CLEF 2006: temporal expressions are better normalized and temporal deictic expressions are resolved to explicit date representations (Sect. 2.1); the coreference module was extended by a fallback strategy for increased robustness (Sect. 2.2); equivalence rules can now introduce *negated information* that must not occur in document sentences used for generating answers (Sect. 2.3); answer candidates are clustered in order to avoid multiple occurrences of one real-world entity in the answers to a list question (Sect. 2.4); and finally a shallow QA system that produces a second answer stream was combined with InSicht (Sect. 2.5).


```

((rule (
  (?r1 ?n1 "übermorgen.1.1") ; temporal adverb referring to the day after tomorrow
  (member ?r1 (ante dur fin strt temp))
→
  ($delete (?r1 ?n1 "übermorgen.1.1"))
  (?r1 ?n1 ?n2)
  (attr ?n2 ?n3) (sub ?n3 "jahr.1.1") (val ?n3 ?n4)
  (card ?n4 (date-year (plus-days ?now 2)))
  (attr ?n2 ?n5) (sub ?n5 "monat.1.1") (val ?n5 ?n6)
  (card ?n6 (date-month (plus-days ?now 2)))
  (attr ?n2 ?n7) (sub ?n7 "tag.1.1") (val ?n7 ?n8)
  (card ?n8 (date-day (plus-days ?now 2))))
(name "deixis.übermorgen"))

```

Fig. 1. A sample rule for resolving temporal deixis. The concepts *jahr.1.1* (*‘year’*), *monat.1.1* (*‘month’*), and *tag.1.1* (*‘day’*) are attributes used in normalized date representations with MultiNet.

2 Improvements over the System for QA@CLEF 2005

2.1 Temporal Expressions

As many factoid questions are related to temporal expressions, InSicht was extended to use not just explicit temporal expressions like *before September 18, 2005* and *in 2001* but also deictic expressions like *today* or *2 years ago*. To this end, around 40 MultiNet [2] rules were developed that resolve temporal deictic expressions into explicit forms. For a wider perspective on *indexicals* (or deictic expressions) in general with their classificatory, relational, and deictic components, see [3].

Before the rules can be applied one must determine for each document the date it was published. (The variable *?now* in our deixis resolution rules contains this value.) The date encoded in the document identifier was taken as a starting point, which was refined (if possible) by looking for a date given at the start of the article text. In 61.4% of the articles, the date could be refined from the text; in the remaining 38.6% (which all came from the FR newspaper subcorpus of the QA@CLEF corpus and not from the other two subcorpora, SDA and SPIEGEL), only the document identifier was used. Currently, the SGML attribute *DATE* (*DT*, *WEEK*) is not used because it is often inaccurate: the FR subcorpus contains only a *WEEK* attribute which corresponds to the Sunday or Monday starting the week. For the SDA subcorpus, the identification method from article text yields the same results as the *DT* attribute.

Fig. 1 shows a deixis resolution rule that InSicht applies in the document processing step. The rules are transformation rules on semantic networks that can introduce new relations and delete existing relations (operation: *\$delete*). Variables start with a question mark; by convention, node (relation) variables are of the form *?n* (*?r*) followed by an integer, e.g. *?n1* and *?r1* in the first premise term of the rule *deixis.übermorgen*. In addition to testing the presence or absence

of a relation and testing the value of a node feature, the premise can also contain some predefined declarative predicates (e.g. the *member* predicate in the example rule). On the right-hand side of a rule, the value of node variables can be stated by constants or by functional expressions (e.g. functions for calculating with dates like *plus-days*, functions for accessing components of a date like *date-day*, and week-aware rounding functions).

In addition to rules that introduce canonical date representations (details below the resolution of days are currently ignored because they seem to be less prominent in questions over newspaper and newswire articles), some preparatory normalization rules were needed. Examples are adjectives that correspond to deictic adverbs, e.g. the adjective *morgig* corresponding to the adverb *morgen*/'*tomorrow*'. The semantic representations of these adjectives are currently flat in the sense that they are similar to the ones of other operational qualities in MultiNet. Therefore, preparatory normalization rules introduce the temporal interpretation of these adjectives, e.g. for *das morgige Treffen*/'*tomorrow's meeting*' a temporal relation TEMP from *treffen.1.1*¹ to *morgen.1.1*. The newly introduced temporal concepts like *morgen.1.1* are then in turn normalized by rules like the one in Fig. 1. Note that the rules can be easily transferred to other languages because they work on semantic representations of the MultiNet formalism. Only some lexical concepts (like the days of the week) must be translated; if one already has a mapping between the involved lexicons, rules can be *translated* automatically for the use in another language.

For some sentences, the deictic expression has been introduced by the prepositional phrase (PP) interpretation of the WOCADI parser. For example, the *vor*-PP in the noun phrase *das Konzert vor 20 Tagen*/'*the concert 20 days ago*' introduces an artificial concept *now.0* (besides a relation to the semantic representation of *20 days*) that needs to be resolved by using the notion of publication date mentioned above.

The ten most successful deixis resolution rules (i.e. the ten rules that fired most often for semantic networks of document sentences) are listed in Table 1. The percentages of rule application for complete parses and for whole texts from the subcorpora in Table 2 reflect different text types. The SDA subcorpus (newswire text) contains short texts with a large percentage of temporal deixis (almost 1 out of 10 sentences). The reference to past days of the week is extremely frequent because the current day of newswire text is very prominent to the reader or hearer. In the daily newspaper subcorpus (FR), the distribution of firing rules is more uniform than in the SDA subcorpus. And finally, the weekly newspaper subcorpus (SPIEGEL or SPI) shows the smallest rule activation rates because a weekly newspaper has longer articles and cannot clearly refer to *yesterday*, certain days of the week, or similar concepts because it is unknown on what day the typical reader of a weekly newspaper will read the article.

A deixis resolution that helped answering question qa06_079 (from QA@CLEF 2006) was triggered by a *vor*-PP in document SDA.951109.0236:

¹ A lexical concept identifier in our lexicon consists of the lemma, a numerical homograph identifier, and a numerical reading identifier.

Table 1. Description of rules for resolving temporal deixis

Deixis rule	Natural language example
weekday.past	<i>The students met last Monday.</i>
today	<i>The book is published today.</i>
weekday.nonpast	<i>The group will discuss on Friday.</i>
monthname.past	<i>The law was passed in August.</i>
year+opminus	<i>the match 20 years ago</i>
tomorrow	<i>The president will arrive tomorrow.</i>
monthname.nonpast	<i>The team will win in December.</i>
today's	<i>today's newspaper</i>
year+prop.past	<i>Most companies were successful in the past year.</i>
yesterday	<i>The peace treaty was signed yesterday.</i>

- (1) In welchem Jahr starb Charles de Gaulle?
'*In which year did Charles de Gaulle die?*'
- (2) Frankreichs Staatschef Jacques Chirac hat die Verdienste des vor 25 Jahren gestorbenen Generals und Staatsmannes Charles de Gaulle gewürdigt. (SDA.951109.0236)
'*France's chief of state Jacques Chirac acknowledged the merits of general and statesman Charles de Gaulle, who died 25 years ago.*'
- (3) 1970

For question qa06_172,² the interpretation of a week day was the key to answering correctly:

- (4) An welchen[sic] Tag haben Jordanien und Israel den Friedensvertrag unterzeichnet?
'*On which day did Jordan and Israel sign the peace treaty?*'
- (5) Israel und Jordanien haben am Mittwoch im Grenzgebiet zwischen beiden Ländern einen Friedensvertrag unterzeichnet. (SDA.941026.0140)
'*Israel and Jordan signed a peace treaty in the border area between both countries on Wednesday.*'
- (6) am 26.10.1994 ('*on 26th October 1994*')

For related work, see for example [4] which briefly describes a rule-based translation from the output of an English semantic parser to an OWL-Time representation for temporal expressions, especially temporal aggregates. [5] covers similar cases like the deixis resolution rules in InSicht but the system looks for lexical and syntactic clues based on part-of-speech tags and does not build on semantic representations derived from full syntactico-semantic parsing like InSicht does. [6] describes a system for German financial news that extracts temporal information based on a part-of-speech tagger and specialized finite state trans-

² The official question contains a grammatical error, which InSicht's parser ignored.

Table 2. Statistics on rule application for temporal deixis

Deixis rule	Application (% of semantic networks)				Application (% of texts)			
	FR	SDA	SPI	all	FR	SDA	SPI	all
weekday.past	1.41	6.49	0.07	3.58	10.99	39.62	0.93	25.00
today	0.80	0.58	0.84	0.70	6.13	3.82	10.10	5.16
weekday.nonpast	0.49	1.00	0.05	0.69	4.05	7.06	0.74	5.41
monthname.past	0.18	0.49	0.16	0.32	1.57	3.26	2.35	2.47
year+opminus	0.23	0.23	0.28	0.23	2.07	1.68	4.01	1.97
tomorrow	0.12	0.37	0.04	0.22	1.03	2.66	0.48	1.83
monthname.nonpast	0.17	0.25	0.10	0.20	1.52	1.81	1.50	1.67
today's	0.20	0.15	0.07	0.16	1.80	1.07	1.11	1.40
year+prop.past	0.13	0.15	0.10	0.14	1.14	1.10	1.38	1.13
yesterday	0.25	0.01	0.02	0.12	2.05	0.03	0.26	0.94
<i>one or more rules</i>	3.79	9.60	1.88	6.25	26.83	52.92	21.88	39.81

ducers. In contrast to [5], InSicht and [6] also take into account the semantics of PPs (see [7] for details on the underlying PP interpretation in InSicht's parser).

2.2 Robust Coreference Resolution

As only 45% of the texts received a partition from the coreference resolution module CORUDIS [8,1] in QA@CLEF 2005,³ a fallback strategy was added to InSicht. If no partition of mentions (markables) can be found by CORUDIS, a *fallback partition* of mentions is calculated as follows. All pronouns are resolved to their most likely antecedents (as estimated by CORUDIS); all other mentions are ignored, i.e. each of these mentions ends up in a singleton set, which is an element of the fallback partition.

2.3 Negated Information for Network Matching

The open world assumption (OWA) in semantic network representations can be problematic if the *query expansion* step produces semantic networks that are similar to the original question network but less specific. For example, InSicht employs a rule named *drop_first_name* that drops the first name of a person if the last name is also in the question (see Fig. 2). Without additional measures, the resulting query network will also match a document network containing a first name, even though this first name might differ from the one in the original question. Therefore transformation rules can introduce negated information. In the given example rule, the conclusion specifies via a negated formula that the question network node *?n1* of a person whose first name is deleted by the rule must not contain any first name (*vorname.1.1*) in document networks.

³ This percentage was lower than possible because of parameter settings that allow efficient coreference resolution for the 277,000 texts (without duplicates).

```

(rule (
  (attr ?n1 ?n2) (sub ?n2 "nachname.1.1") (val ?n2 ?) ; last name
  (attr ?n1 ?n3) (sub ?n3 "vorname.1.1") (val ?n3 ?n4) ; first name
→
  ($delete (attr ?n1 ?n3)) ($delete (sub ?n3 "vorname.1.1")) ($delete (val ?n3 ?n4))
  (not (attr ?n1 ?n5) (sub ?n5 "vorname.1.1")))
(name "drop_first_name")
(quality 0.6))

```

Fig. 2. A transformation rule that introduces negated information

2.4 Answer Clustering

The QA system InSicht did not exploit the relationship among answer candidates in previous years; it only combined identical answer candidates into one answer candidate with an increased frequency score that influenced answer selection. But as InSicht was extended to answer list questions, clustering became unavoidable because otherwise the system could return many list items (meant as different entities) which all refer to the same entity. For example, if one asks for the president of a country, one could receive one person in many forms: with all first names, with some first names, with initials, only the last name, etc.

Non-identical answer candidates a and b are clustered together (assuming that they refer to the same entity) if the words in phrase a can be aligned with the ones in b , possibly with additional words or word parts. For example, the answers *Peter Schmidt* and *P. Schmidt* are in the same cluster. This surface definition of a clustering condition will be replaced by a semantically based definition for the semantic networks corresponding to the answer candidates.

2.5 Shallow QA Subsystem

InSicht's parser, WOCADI, produces a partial (and not a complete) semantic network for about 50% of all sentences in the German QA@CLEF corpus. For example, WOCADI can rarely analyze sentences containing grammatical errors, spelling errors, or conflated sentence parts originating from erroneous document preprocessing. Thus, InSicht will not be able to find many of those answers appearing in malformed sentences. For QA@CLEF 2006, we experimented with a shallow QA approach (see [9] for details) which can be seen as producing an additional *answer stream* (see [10] for a QA system with many answer streams).

We applied a sentence boundary detector and a tokenizer to the corpus and split all documents into single sentences. Question-answer pairs for QA@CLEF in previous years were extracted from the MultiEight corpus and augmented manually yielding 732 question-answer pairs. These pairs were fed into a standard IR system to determine candidate sentences containing an answer. The answer candidates were then processed as follows: Words from closed categories were annotated with part-of-speech information, keywords in the query were replaced with symbols representing their type (name, upper-case word, lower-case

word, or number) and the answer string was replaced with an answer variable representing one or more words. We then formed several patterns from this substitution, including a maximum of five tokens left and right of the variable.

In this shallow approach, QA consists of five steps: 1. Perform relevance-ranked IR on the corpus and determine the top 250 ranked sentences. 2. Apply pattern matching on these sentences with the answer patterns extracted. 3. If a match is found, the answer variable contains the string representing the answer. 4. Answer candidates are validated using a simple heuristic: all answers starting with punctuation marks or containing words from a closed word category only are eliminated. 5. Answers are ranked by cumulative frequency and the most frequent answer is chosen.

3 Run Description and Evaluation

The current QA system has been evaluated on the German questions from QA@CLEF 2006. The shallow approach described in Sect. 2.5 found 17 answers in total for the 198 questions assessed for QA@CLEF 2006. Thirteen of them were also found by the deep approach (run FUHA061dede in Table 3), but three of them were new and correct and one was new and inexact (because it omitted a phrase-final acronym). This was our first attempt to combine the deep processing with a shallower method and it leaves many chances for further improvements. However, it already shows that a combination of processing methods will further improve the performance for the IRSAW system, a QA framework which will encompass the QA system InSicht. The combination of the deep approach and the shallow approach corresponds to run FUHA062dede in Table 3.

The result rows preceded by a minus sign (–) in Table 3 are from an ablation study: each extension from Sect. 2 was omitted in an evaluation run in order to see the impact on system performance. All extensions show positive effects; similar improvements occur for the questions from QA@CLEF 2004 and 2005. But their statistical significance will probably appear only when evaluating on many more questions.

Table 3. Results for the German question set from QA@CLEF 2006. W_n (W_{nn}) is the number of wrong NIL answers (wrong non-NIL answers).

Setup and modification	Results					
	#Right	#Uns.	#Inex.	W_{nn}	W_n	K1
InSicht (= FUHA061dede)	62	4	0	3	129	0.1799
– deixis resolution (Sect. 2.1)	62	2	1	3	130	
– robust coreference resolution (Sect. 2.2)	61	4	0	3	130	
– negated information (Sect. 2.3)	62	4	0	5	127	
– answer clustering (Sect. 2.4)	60	3	3	3	129	
+ shallow QA (Sect. 2.5, = FUHA062dede)	65	4	1	3	125	0.1895

4 Conclusion

The extensions of the QA system InSicht after QA@CLEF 2005 (temporal deixis resolution, increased robustness of coreference resolution, use of negated information during network matching, clustering of answers) improved system results. To investigate the statistical significance of these improvements, the tests should be repeated on more questions, especially on more difficult questions.

In the future, more deixis resolution rules need to be written, e.g. for resolving names of fixed and floating holidays (*during Christmas* and *50 days after Easter Sunday*), season names (*before winter*), and vague temporal expressions (*at the end of this year*). Temporal expressions are currently linked to individual sentences. Propagating temporal expressions from a sentence to parts of its surrounding text (*co-text*; e.g. by event ordering) will make much more temporal information available.

References

1. Hartrumpf, S.: Extending knowledge and deepening linguistic processing for the question answering system InSicht. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 361–369. Springer, Heidelberg (2006)
2. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Heidelberg (2006)
3. Nunberg, G.: Indexicality and deixis. *Linguistics and Philosophy* 16, 1–43 (1993)
4. Pan, F., Hobbs, J.R.: Temporal aggregates in OWL-Time. In: Proceedings of the 18th Florida Artificial Intelligence Conference (FLAIRS-05), pp. 560–565 (2005)
5. Mani, I., Wilson, G.: Robust temporal processing of news. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, pp. 69–76 (2000)
6. Schilder, F., Habel, C.: From temporal expressions to temporal information: Semantic tagging of news messages. In: Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, Toulouse, France, pp. 65–72 (2001)
7. Hartrumpf, S., Helbig, H., Osswald, R.: Semantic interpretation of prepositions for NLP applications. In: Proceedings of the Third ACL-SIGSEM Workshop on Prepositions, Trento, Italy, pp. 29–36 (2006)
8. Hartrumpf, S.: Coreference resolution with syntactico-semantic rules and corpus statistics. In: Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001), Toulouse, France, pp. 137–144 (2001)
9. Leveling, J.: On the role of information retrieval in the question answering system IRSAW. In: Proceedings of the LWA 2006 (Learning, Knowledge, and Adaptability), Workshop Information Retrieval, Universität Hildesheim, Hildesheim, Germany, pp. 119–125 (2006)
10. Ahn, D., Jijkoun, V., Müller, K., de Rijke, M., Schlobach, S., Mishne, G.: Making stone soup: Evaluating a recall-oriented multi-stream question answering system for Dutch. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 423–434. Springer, Heidelberg (2005)

Relevance Measures for Question Answering, The LIA at QA@CLEF-2006

Laurent Gillard, Laurianne Sitbon, Eric Blaudez,
Patrice Bellot, and Marc El-Bèze

LIA, Avignon University Science Laboratory,
339 ch. des Meinajaries, BP 1228, F-84911 Avignon Cedex 9, France
{laurent.gillard,laurianne.sitbon,eric.blaudez,
patrice.bellot,marc.elbeze}@univ-avignon.fr
<http://www.lia.univ-avignon.fr>

Abstract. This article presents the first participation of the Avignon University Science Laboratory (LIA) to the Cross Language Evaluation Forum (CLEF). LIA participated to the monolingual Question Answering (QA) track dedicated to the French language, and to the cross-lingual English to French QA track. After a description of the strategies used by our system to answer questions, an analysis of the results obtained is done in order to discuss and outline possible improvements. Two strategies were used: for factoid questions, answer candidates are identified as Named Entities and selected by using key terms density measures (“compactness”); on the other hand, for definition questions, our approach is based on detection of frequent appositive informational nuggets appearing near the focus.

Keywords: Question answering, Questions beyond factoids, Definition questions, Key Terms density metrics, Compactness, Multilingual question answering.

1 Introduction

Question Answering (QA) systems aim at retrieving precise answers to questions expressed in natural language rather than list of documents that may contain an answer. They have been particularly studied since 1999 and the first large scale QA evaluation campaign held as a track of the Text REtrieval Conference [1]. The Cross Language Evaluation Forum (CLEF) has been studying multilingual issues of QA and has been providing an evaluation platform for QA dedicated to many languages since 2003.

It was the first participation of the Avignon University Science Laboratory (LIA) to CLEF (we already participated to monolingual English, TREC-11 [6], and monolingual French, the EQueR Technolangue QA campaign [8]). This year, LIA participation was devoted to two tracks: monolingual French and English to French. Both submissions are based on the same system architecture that was initially developped for our participation to the EQueR campaign[2]. Moreover,

CLEF-2006 allowed us to evaluate some improvements: English question analysis module, integration of Lucene as a search engine, a module to handle definition questions, and another one to experiment a re-ranking mechanism by using redundancy for answer candidates. For both tracks, two runs were submitted, the second run for each language used a redundancy re-ranking module.

Our QA system (called SQuaLIA) is built according to the typical QA system (QAS) architecture, and involves pipelined main components. A Question Analysis (described in section 2) is first done to extract the semantic type(s) of the expected answer(s) and key terms but also to decide what strategy must be followed : factoid or definition one. For factoid questions (section 3), SQuaLIA performs consecutive steps: Document Retrieval (section 3.2) to restrict the amount of processed data for next components, Passage Retrieval (section 3.3) to choose the best answering passages from documents, and lastly, Answer Extraction to determine the best answer candidate(s) drawn from the previously selected passages. This answer extraction is mainly performed by using a density (section 3.4) of the key terms appearing around an answer candidate (bounded as a Named Entity, NE, section 3.1), but may also involve knowledge bases (section 3.4). The definition strategy (section 4) uses frequent appositive nuggets of information appearing near the focus. We also experimented briefly with redundancy to re-rank answer candidates but this module had some bugs, and did not perform as expected. In last sections, results of our participation will be presented and discussed (section 5) with perspectives for future improvements.

2 Question Analysis

Firstly, definition questions are recognized with a simple pattern matching process, which also extracts the focus of the question. All questions which do not fit definition patterns are considered as factual questions, including list questions (and consequently, list questions were wrongly answered as factoid questions with only one instance instead of a set of instances).

The analysis of factual questions is made of two main independent steps which are: question classification and key terms extraction. For the questions in English, the key terms extraction step is done after a translation of the question by using Systran translator¹.

Expected Answer Types Classification: The answer expected for a factoid question is considered to be a Named Entity (NE) based upon Sekine's hierarchy [3]. Thus, questions are classified according to this hierarchy. This determination is done with patterns and rules for questions in French; and, complemented, but only for English questions, with semantic decision trees (a machine learning approach introduced by [4] and [5]).

Semantic decision trees are adapted as described in [6] to automatically extract generic patterns from questions in a learning corpus. The corpus *CLEF multinline* composed of 900 questions was used for learning. Since all questions

¹ By means of Google Translator: http://www.google.fr/language_tools

Table 1. Evaluation of the categorization process on CLEF-2006 questions

Language	Correct	Wrong	Unknown
French	86%	2%	12%
English	78%	13%	5%

are available in both English and French languages, French questions were automatically tagged according to the French rules based module, then, checked and paired manually with their English version to be used as a learning corpus. The learning features are words, part-of-speech tags², lemmas, and length of the question (stop words are taken into account).

Table 1 shows the results of a manual post-evaluation of the categorization of CLEF-2006 questions (first and second columns), but also the percentage of questions tagged as *Unknown* by categorizer (third column). Wrong or *Unknown* tags prevent the downstream components from extracting answers. In French, only 8 of the 19 *Unknown* tagged questions were classifiable manually (as being *Person* name or *Company* name, others are difficult to map to Sekine’s hierarchy). Due to its “strict” pipelined architecture, the percentage of Correct tags fixed the maximum score our QAS can achieve. Moreover, some of the correct classification in English are actually more generic than they could be, for example *President* is tagged as *Person*.

After assigning questions to classes from the hierarchy, a mapping between these classes and available Named Entities (NE) is done. However, the question hierarchy is much more exhaustive than the named entities that our NE system is able to recognize.

Key Terms Extraction: *Key Terms* are important objects like words, named entities, or noun phrases extracted from questions. These key terms are expected to appear near the answer. Actually, our CLEF 2006 experiments used only lemmatized forms of nouns, adjectives and verbs as key terms. But, noun phrases may also help in the ranking of passages. For example, in question Q0178, “*world champion*” is more significant than “*world*” and “*champion*” taken separately. Such noun phrases are extracted in French with the help of SxPipe deep parser described in [9], Named Entities are extracted as described in section 3.1. For English questions, these extractions are done on their French equivalent obtained after initial translation.

3 Answering Factoid Questions

Answering factoid questions was done by using the strategy we used for our participation to EQueR. It mainly relies on density measures to select answers which are sought as Named Entities (NE) paired with expected answer types (EAT). Two new modules were developed for this CLEF-QA’s participation: Lucene

² Part-of-speech tags and lemmas are obtained with the help of TreeTagger [7].

was integrated as the Document Retrieval (DR) search engine (for EQueR as for TREC, Topdoc lists were available, and so, this DR step was facultative) and we also experimented with redundancy to re-rank candidate answers.

3.1 Named Entity Recognition

Named Entities detection is one of the key elements as each answer that will be provided must first be located (and bounded) as a semantic information nugget in the form of a Named Entity. Our named entities hierarchy was created from and is a subset of the Sekine's taxonomy [3]. NE detections are done by using automata: most of them are implemented by using GATE platform, but also by direct mapping of many gazetteers gathered from the Web.

3.2 Document Retrieval

The Document Retrieval (DR) step aims at identifying documents that are likely to contain an answer to the question posed and, thus, restrict search space for next steps. Indexation and retrieval were done by using Lucene search engine with its default similarity scoring³. For each document of the collection, only lemmas are indexed after a stop-listing based upon their TreeTagger's part-of-speech. Disjunctive queries are formulated from the lemmas of the question key terms. At retrieval time, only the first 30 documents (at most) returned by Lucene were considered and passed to the Passage Retrieval component.

3.3 Passage Retrieval

Our CLEF passage retrieval component considers a question as a set of several kinds of items: lemmas, Named Entity tags, and expected answer types.

First, a density score s is computed for each occurrence o_w of each item w in a given document d . This score measures how far are the items of the question from the other items of the document. This process focuses on areas where the items of the question are most frequent. It takes into account the number of different items $|w|$ in the question, the number of question items $|w, d|$ occurring in the document d and a distance $\mu(o_w)$ that represents the average number of items from o_w to the other items in d (in case of multiple occurrences of an item, only the nearest occurrence to o_w is considered).

Let $s(o_w, d)$ be the density score of o_w in document d :

$$s(o_w, d) = \frac{\log[\mu(o_w) + (|w| - |w, d|) \cdot p]}{|w|} \quad (1)$$

where p is an empirically fixed penalty. The score of each sentence S is the maximum density score of the question items it contains:

$$s(S, d) = \max_{o_w \in S} s(o_w, d) \quad (2)$$

³ Lucene version 1.4.3, using TermVector and similarity: <http://lucene.apache.org/java/docs/api/org/apache/lucene/search/Similarity.html>

Passages are composed of a maximum of three sentences: the sentence S itself, the previous and following sentences (if they exit). The first (and at most) 1000 best passages are then considered by the Answer Extraction component.

3.4 Answer Extraction

Answer extraction by using a density metric. To choose the best answer to provide to a question, another density score is calculated inside the previously selected passages for each answer candidate (a Named Entity) adequately paired with an expected answer type for current question. This density score (called *compactness*) is centered on each candidate and involves extracted key terms (let $QSet$ be this set) from the question during the question analysis step.

The assumption behind our *compactness* score is that the best candidate answer is closely surrounded by the important words (actually extracted key terms) of the question. Every word not appearing in the question can disturb the relation between a candidate answer and its responsiveness to a question. Moreover, in QA, term frequencies are not as useful as for Document Retrieval: an answer word can appear only once, and it is not guaranteed that words of the question will be repeated in the passage, particularly in the sentence containing the answer. A score improvement can come from incorporating an inverse document frequency or linguistic features for non $QSet$ encountered words to further take into account any variation of closeness.

For each *answer candidate* AC_i , compactness score is computed as follow:

$$compactness(AC_i) = \frac{\sum_{y \in QSet} p_{y_n, AC_i}}{|QSet|} \quad (3)$$

with y_n the nearest occurrence key term y from AC_i and:

$$p_{y_n, AC_i} = \frac{|W|}{2R + 1} \quad (4)$$

$$R = distance(y_n, AC_i) \quad (5)$$

$$W = \{z | z \in QSet, distance(z, AC_i) \leq R\} \quad (6)$$

For both runs submitted, only unigrams were considered for inclusion in $QSet$. All answers candidates are then ranked by their score, best scoring answer is provided as final answer.

Answer extraction by using knowledge databases. For some questions, answers are *quite* invariable over time like, for example, questions asking about the capital of a country, the authors of book, etc.

To answer these questions one may use (static) knowledge databases (KDB). We had built such KDB for our participation to TREC-11 [6], and translated them to a French equivalent for our participation to EQueR (they were particularly tuned on CLEF-2004 questions). For CLEF-2006, we used the same unchanged KDB module. It provided answer patterns, which were used to mined

Table 2. Knowledge databases coverage for CLEF 2004-06

runs	2004	2005	2006 FR-FR
Q covered (/200)	51	26	24 (10)
Passages matching pattern	48	<i>ne</i>	11
Correct answers (C)	38-40	<i>ne</i>	11

passages from Passage Retrieval step. Table 2 shows the coverage of these databases for CLEF from 2004 to 2006 (*ne* stands for “not evaluated”), the number of passages coming from PR matching a KDB pattern, and number of right supported answers. Also, for CLEF 2006 experiments, 10 questions covered by this module were finally handled directly by the new definition module). EN-FR results were lower due to translation errors (9C or 7C).

4 Answering Definition Questions

For our participation to EQueR [2], most of the definition questions were unanswered by the strategy described in the previous section. Indeed, it mainly relies on the detection of an entity paired with a question type, but, for definition questions, such pairing to an entity is far more difficult: the expected answer can be anything qualifying the object to define (even if, when asking about a person definition, the answer sought is often its main occupation - which may constitute a common Named Entity - it can still be tricky to recognize all possible occupations or reasons why someone maybe “famous”). Also, one can notice that while for TREC-QA campaign all vital nuggets should be retrieved, for the EQueR and CLEF exercises, retrieving only - one - vital was sufficient.

So, the problem we wanted to address was to find one of the best definitions available inside the corpora. Therefore, we developed a simple approach based on appositive and redundancy to deal with these questions:

- the focus to define is extracted from the question and is used to filter and keep all sentences of the corpora containing it.
- then, appositives nuggets such as $< focus , (comma) nugget , (comma) >$, $< nugget , (comma) focus , (comma) >$, or $< definite article (“le” or “la”) nugget focus >$ are sought.
- the nuggets are divided in two sets: the first, and preferred one, contains nuggets that can be mapped, by using their TreeTagger part-of-speech tags, on a minimal noun phrase structure while the second set contains all others (it can be seen as a kind of “last chance” definition set).
- both sets are ordered by their nuggets frequency in the corpora. But the “noun phrase set” is also ordered by taking into account its head noun frequency, the overlapping count of nouns inside it among the most frequent nouns appearing inside all the nuggets retrieved and its length. All these frequency measures are aimed to choose what is expected to be “the most common informative definition”.

- best nugget of the first set or of the second, if first is empty, or by default NIL, is provided as final answer.

This strategy was also used to answer to the acronyms/expanded acronyms definitions, as the same syntactic punctuation clues can be used for these questions such as the frequent *< acronym ((opening parenthesis) expanded acronym) (closing parenthesis) >* or *< expanded acronym ((opening parenthesis) acronym) (closing parenthesis) >* - and it performed very well as all answers of this year set were retrieved (even if, due to a bug in the matching process between answer and justification, the “TDRS”/Q0145 wasn’t finally answered, while the correct answer was found).

5 Results and Analysis

On the 200 test questions, and for our best run (FR-FR1), 93 right answers (89 + 4 lists) were correct. Table 3 presents the breakdown according to Factual, Definition and Temporally restricted questions of our two basic runs (i.e. without re-ranking, our second runs, which contained errors).

Table 3. CLEF-QA 2006 breakdown for our basic FR-FR and EN-FR runs, derived from official evaluation

runs	Correct Answers				NIL Answers		others	
	Fact.	Def.	Temp.	Total	Correct	answered	ineXact	Unsupp.
FR-FR1	44	33	12	89	2	33	7	2
EN-FR1	31	28	8	67	5	35	7	2

Results analysis: because our QAS is composed of pipelined components, early silences or errors in the chain lead to incorrect or no final answers. As shown in Table 4, and without verifying whether a correct answer is present in the documents, early silences coming from Information Retrieval (IR) step (first row), missing or wrong pairing between Expected Answer Types (EAT) and Named Entities (NE) (second row), or the fact that EAT is wrongly assigned (third row) may all cause a final error. Thus for FR-FR run, our IR step prevents from finding any correct answer for 2% of questions, errors between NE and EAT is up 17%, and not searching for correct EAT raises this percent to 19%. Similarly, For EN-FR, and due to these earliest errors, no correct answer might be mined for 28% of questions.

Knowledge Databases: The knowledge databases best contribution added 11 correct answers. However, without using KDB, 5 of these would have been also found. Final best KBD contribution added 6 correct answers.

IneXact answers: If we examine all the inexact answers from all our 4 runs, 8 of 20 where *Date* and 3 of 20 were *Number of people*. For *Date*, 6 of 8 were *correct*

Table 4. Early causes of silence or incorrect answers

	FR-FR1 (basic run)					EN-FR1 (basic run)				
	Fact.	Temp.	List	All		Fact.	Temp.	List	All	
#Q	109	43	6	158	% Loss	109	43	6	158	% Loss
IR Q→D	108	41	6	155	2%	109	43	6	158	0%
NE↔EAT	95	30	6	131	17%	96	37	5	138	13%
EAT is OK	94	29	5	128	19%	83	28	3	114	28%
% Loss	14%	33%	17%			25%	35%	50%		

date answers but missed the year because it did not appear inside supporting sentence. But, for 3 of 6, missing year could be easily located elsewhere in the document by matching pattern “19[0-9]/[0-9]” or was the publication year for the 3 others (and document did not cite other year). So, simple detection rules, with default value to the year of the document, could have helped to answer these questions. Lastly, one other *Date* question judged as inexact is actually wrong: Australia will probably never be a European Union member (Q0180). For *Number of people* questions, all 3 were missed due to bad detection of NE boundaries, but 2 because of the lack of “près d[e]’/around”.

Temporally restricted factoid questions: were handled as non restricted factoid ones, that is, dates appearing inside questions were considered just like another key term. Correct answers were found for at most 12 of these questions.

Definition questions: Results for the definition questions were quite good among the 41 definitions, our best run (FR-FR) answered 32 correct plus 1 NIL correct (“Linux”/Q0003, the word Linux never appear inside corpora). Among the 5 incorrect answers provided: 2 of them were due to development errors and should have been answered correctly; 1 expected a NIL answer (“T-shirt”/Q0144), and cannot be answered by our approach (which tends to always provide an answer unless the object to define did not appear inside corpora, it misses some confidence or validation measure to discard very improbable answer); and the last 2 needed more complex resolutions: answers were before or after a relative clause but they were also separated from the focus by using more than one punctuation (a dependency tree could have help to provide them). For the 3 inexact definition answers provided: “Boris Becker”/Q0090 is hard to define without using any anaphora resolution, the string “Boris Becker” never co-occurs inside a sentence with a pattern “Tennis” (“joueur/player” or “man”) or any “vainqueur/winner”. The 2 others also need to extract a subpart of a relative clause, but the correct corresponding passages were located in the top five definitions selected.

Results for the EN-FR runs were lower (best one was 27 + 1 NIL/41) due to translation errors.

List questions: Our QAS wrongly classified the 6 list questions⁴ as factoid questions and answered with only one answer by question. For our basic FR-FR run, we provided correct answers for 4 questions, and for 3 of these, the justification contained all other needed correct answers. For our basic EN-FR run, only 1 correct answer was provided but the justification contained all the other correct instances, the second run found 2 correct answers but justifications did not contain other needed instances.

6 Conclusion and Future Works

We have presented the Question Answering system used for our first participation to the QA@CLEF-2006 Evaluation. We participated in monolingual French (FR-FR), and cross-lingual English to French (EN-FR) tracks. Results obtained were acceptable with an accuracy of 46% for FR-FR, and 35% for EN-FR.

However, performance can also easily be improved at every step as shown in the detailed analysis of our results. For example, our broken redundancy criterion, by changing order among the top five answers, allowed us to answer some questions that were previously wrongly answered.

Inadequate Named Entities (NE) were also a bottleneck in the processing chain, and many of them should still be added (or improved). To compensate this known weakness, we tried to include in our system a simple reformulation module, notably for the “*What/Which (something)*” questions. It would have been in charge of verifying answers already extracted by the factoid extraction strategy and, particularly, of completing it when NE was unknown. But our preliminary experiments showed us that such question rewriting was tricky. To our opinion, questions were (intentionally or not) formulated such that it is difficult by changing words order to match an answering sentence (synonyms might help here). Another problem is, when using the Web to match such reformulation, results obtained are noisy: we experiment it for presidents or events (Olympic Games) but CLEF corpora deals with years 1994 and 1995 whereas Web emphasizes on recent era.

With an accuracy of 76%, performance on definition questions was quite good even if it could still benefit from anaphora resolution. Moreover, our methodology is mostly language independent as it does not involve any language knowledge (other than detecting appositive mark up, a similar approach was used by [10] to answer Spanish definition questions). We are interested in improving it in many ways: first, to locate more difficult definitions (as the one we missed); but also to choose better qualitative definitions (by combining better statistics), and finally to synthesize these qualitative definitions. For example, “*Bill Clinton*” was defined as an “*American president*”, “*Democrat president*”, “*democrat*”, “*president*”, he can be best defined as an “*American democrat president*”. “*Airbus*” was defined as “*European consortium*”, or “*Aeronautic consortium*”, so defining it as a “*European Aeronautic consortium*” could be even better.

⁴ (4 questions were misclassified by the CLEF staff and are not actual list questions: Q0133, Q0195, Q099, Q0200).

Concerning date and temporally restricted questions, we could not achieve the development of a complete processing chain, so this is a future work (actually, by the time of the CLEF evaluation, we only have built a module to extract time constraints from the question). We also need to complete our post-processing to select or to synthesize complete date (with the difficulties one can expect for relative dates).

Lastly, our QAS still misses confidence scores after each pipelined components to be able to do: some constraint relaxations (for example on key terms), any loop back or simply to decide when a NIL answer must be provided.

References

1. Voorhees, E.M.: Chapter 10. Question Answering in TREC. In: Voorhees, E.M., Harman, D. (eds.) *TREC Experiment and Evaluation in Information Retrieval*, pp. 233–257. MIT Press, Cambridge, Massachusetts (2005)
2. Gillard, L., Bellot, P., El-Bèze, M.: *Le LIA à EQueR* (campagne Technolanguage des systèmes Questions-réponses) In: *proceedings of TALN-Recital*, vol. 2, pp. 81–84. Dourdan, France (2005)
3. Sekine, S., Sudo, K., Nobata, C.: Extended Named Entity Hierarchy. In: *proceedings of The Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, pp. 1818–1824 (2002)
4. Kuhn, R., De Mori, R.: The Application of Semantic Classification Trees to Natural Language Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17, 449–460 (1995)
5. Béchet, F., Nasr, A., Genet, F.: Tagging Unknown Proper Names Using Decision Trees. In: *proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong-Kong, China, pp. 77–84 (2000)
6. Bellot, P., Crestan, E., El-Bèze, M., Gillard, L., de Loupy, C.: Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question-Answering Track. In: *proceedings of The Eleventh Text REtrieval Conference (TREC 2002)* NIST Special Publication, pp. 500–251 (2003)
7. Schmid, H.: Probablistic Part-of-Speech Tagging Using Decision Trees. In: *proceedings of The First International Conference on New Methods in Natural Language Processing (NemLap-94)*, Manchester, U.K, pp. 44–49 (1994)
8. Ayache, C., Grau, B., Vilnat, A.: EQueR: the French Evaluation Campaign of Questions Answering Systems. In: *proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy (2006)
9. Sagot, B., Boullier, P.: From Raw Corpus to Word Lattices: Robust Pre-parsing Processing with SxPipe. *Special Issue on Human Language Technologies as a Challenge for Computer Science and Linguistics, Archives of Control Sciences* 15, 653–662 (2005)
10. Ferrés, D., Kanaan, S., González, E., Ageno, A., Rodríguez, H., Turmo, J.: The TALP-QA System for Spanish at CLEF-2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)

Monolingual and Cross-Lingual QA Using AliQAn and BRILI Systems for CLEF 2006*

S. Ferrández¹, P. López-Moreno¹, S. Roger^{1,2}, A. Ferrández¹, J. Peral¹,
X. Alvarado¹, E. Noguera¹, and F. Llopis¹

¹ Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain

² Natural Language Processing Group
Department of Computing Sciences
University of Comahue, Argentina

{sferrandez,sroger,antonio,jperal,elisa,llopis}@dlsi.ua.es,
P.Lopez@ua.es

Abstract. A previous version of AliQAn participated in the CLEF 2005 Spanish Monolingual Question Answering task. For this year's run, to the system are added new and representative patterns in question analysis and extraction of the answer. A new ontology of question types has been included. The inexact questions have been improved. The information retrieval engine has been modified considering only the detected keywords from the question analysis module. Besides, many PoS Tagger and SUPAR errors have been solved and finally, dictionaries about cities and countries have been incorporated. To deal with Cross-Lingual tasks, we employ the BRILI system. The achieved results are overall accuracy of 37.89% for monolingual and 21.58% for bilingual tasks.

1 Introduction

This paper outlines the participation of AliQAn and BRILI (Spanish acronym for “Question Answering using Inter Lingual Index Module”) systems [2] at CLEF 2006¹. AliQAn [4] participated in CLEF 2005, and this year a new version of our monolingual system is proposed. In addition, the BRILI system [1] is presented to carry out the Cross-Lingual (CL) tasks.

The next section shows the structure, the functionality and main improvements of AliQAn. Section 3 describes our participation in the CL English-Spanish task with the BRILI system. Afterwards, the achieved results are presented in section 4. Finally, we conclude and discuss further work directions.

* This research has been partially funded by the Spanish Government under project CICYT number TIC2003-07158-C04-01 by the Valencia Government under project number GV06/161 and by the University of Comahue under the project 04/E062.

¹ <http://www.clef-campaign.org/2006.html>

2 System Description

This section shows the architecture (see figure 1), new characteristics and improvements of AliQAn with respect to our participation in CLEF 2005 [4].

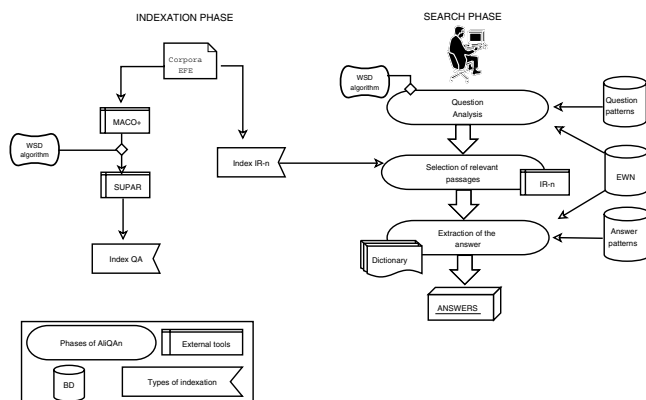


Fig. 1. AliQAn system architecture

New characteristics and improvements have been introduced into AliQAn. Next, these new contributions are detailed:

- In order to correct some of the wrong outputs produced by the PoS tagger and SUPAR systems (syntactic analyzer), a new set of post-processing rules have been introduced into the system. For example, in the case of the Pos Tagger when the word “*El*” is labeled as *proper name*, a rule replaces this tag by the tag *determiner*. Besides, for this participation, SUPAR has allowed the coordinated nominal phrases in the appositions.
- In order to detect the expected answer type in the question analysis phase, a new subset of syntactic patterns has been introduced. Our taxonomy has been expanded in new four types (profession, first name, temporary ephemeride and temporary day).
- A new proposal for Word Sense Disambiguation (WSD) for nouns [3] is applied to improve the system.
- In the current implementation of our systems, the inputs introduced to the passage retrieval system have been only the detected keywords in the question analysis module.
- In the extraction of the answer phase, the patterns applied to identify the answers have been thoroughly studied. Therefore, some of the patterns have been subtracted and new ones have been added. This modifications have enhanced the precision of our system.
- Furthermore, dictionaries of countries and cities have been created and incorporated into the systems. We use them for the questions of type *country* or *city*, because the possible answers may belong to the dictionaries.

3 Cross-Lingual Task

In the case of the CL task, the BRILI system [1] is used. The fundamental characteristic of BRILI is the strategy used for the question processing module in which the Inter Lingual Index (ILI) Module of EuroWordNet (EWN) is used with the aim of reducing the negative effect of question translation on the overall accuracy.

The BRILI system introduces two improvements that alleviate the negative effect produced by machine translation: (1) BRILI considers more than only one translation per word by means of using the different synsets of each word in the ILI module of EWN; (2) unlike the current bilingual English-Spanish QA systems, the question analysis is developed in the original language without any translation.

In order to show the complete process of BRILI, an example of a question at CLEF 2004 is detailed:

- **Question 101 at CLEF 2004:** *What army occupied Haiti?*
- **Syntactic Blocks (SB):**
[NP army] [VP to occupy] [NP Haiti]
- **Type:** group
- **Keywords to be referenced:** army occupy Haiti
 - army $\mapsto$ ejército
 - occupy $\mapsto$ absorber **ocupar** atraer residir vivir colmar rellenar **ocupar** llenar
 - Haiti $\mapsto$ Haití
- **SB to search in Spanish documents:**
[NP ejército] [VP ocupar] [NP Haití]

In the question analysis process, the BRILI system detects the SB needed for the search of the answer and the type of the question which fixes the information that the answer has to satisfy to be a candidate of an answer (proper name, quantity, date, ...).

In the previous example, the system finds more than one Spanish equivalents for one English word. The current strategy employed to solve this handicap consists of assigning more value to the word with the highest frequency. In the case of the previous example, the most valued Spanish word would be “*ocupar*”.

On the other hand, the words that are not in EWN are translated into Spanish using an on-line Spanish Dictionary². Besides, BRILI uses gazetteers of organizations and places in order to translate words that have not linked using ILI. Therefore, in order to decrease the effect of incorrect translation of the proper names, the matches using these words in the search of the answer are carried out using the translated word and the original word of the question. The found matches using the original English word are valued a 20% less.

² <http://www.wordreference.com>

4 Results

Table 1 shows the official results of AliQAn (alia061eses) and BRILI (alia061enes) systems at CLEF 2006. The snippets returned by our systems were shorter with regard to our retrieval passages. Therefore, the unsupported answers have a higher error rate. When some of the answer are not really unsupported.

Table 1. AliQAn and BRILI results at CLEF 2006

	Right	Wrong	Inexact	Unsupported	Overall Accuracy (%)
alia061eses	72	95	15	8	37.89
alia061enes	41	124	9	16	21.58

5 Conclusion and Future Work

For our second participation at QA@CLEF, we presented an improved version of our AliQAn system and a new system for the CL tasks called BRILI. The monolingual performance obtained an improvement of 14.81% and the BRILI system was ranked first at English-Spanish task.

Further work will study the possibility to take into account the multilingual knowledge extracted from Wikipedia³, temporal question treatment and the addition of a Named Entity Recognizer to detect proper names that will not be translated in the question.

References

1. Ferrández, S., Ferrández, A., Roger, S., López-Moreno, P., Peral, J.: BRILI, an English-Spanish Question Answering System. In: Proceedings of the International Multiconference on Computer Science and Information Technology. vol. 1, pp. 23–29 (November 2006) ISSN 1896-7094
2. Ferrández, S., López-Moreno, P., Roger, S., Ferrández, A., Peral, J., Alavarado, X., Noguera, E., Llopis, F.: AliQAn and BRILI QA System at CLEF-2006. In: Workshop of Cross-Language Evaluation Forum (CLEF) (2006)
3. Ferrández, S., Roger, S., Ferrández, A., Aguilar, A., López-Moreno, P.: A new proposal of Word Sense Disambiguation for nouns on a Question Answering System. In: Gelbulkh, A. (ed.) Advances in Natural Language Processing. Research in Computing Science. vol. 18, pp. 83–92 (February 2006) ISSN: 1665-9899
4. Roger, S., Ferrández, S., Ferrández, A., Peral, J., Llopis, F., Aguilar, A., Tomás, D.: AliQAn, Spanish QA System at CLEF-2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 457–466. Springer, Heidelberg (2006)

³ <http://www.wikipedia.org/> - A Web-based free-content multilingual encyclopedia project

The Bilingual System MUSCLEF at QA@CLEF 2006

Brigitte Grau, Anne-Laure Ligozat, Isabelle Robba, Anne Vilnat
Michael Bagur, and Kevin Séjourné

LIR group, LIMSI-CNRS, BP 133 F-91403 Orsay Cedex, France

`firstName.name@limsi.fr`

<http://www.limsi.fr/Scientifique/lir/>

Abstract. This paper presents our bilingual question answering system MUSCLEF. We underline the difficulties encountered when shifting from a mono to a cross-lingual system, then we focus on the evaluation of three modules of MUSCLEF: question analysis, answer extraction and answer fusion. We finally present how we re-used different modules of MUSCLEF to participate in AVE (Answer Validation Exercise).

Keywords: Question answering, evaluation, multiword expressions.

1 Introduction

This paper presents our cross-lingual question answering system, called MUSCLEF, that participated in the QA@CLEF 2006 French-English cross-language task for which two runs were submitted. As for the past two years, we used two strategies: the first one consists in translating only a set of terms selected by the question analysis module, this strategy being implemented in a system called MUSQAT; the second one consists in translating the whole question and then applying our monolingual system named QALC. In the previous CLEF campaigns, most of the systems used exclusively question translation. This year, several systems used either term translation ([1], [2]) or an hybrid approach ([3], [4]) consisting in question translation plus term translation. Approaches based on term translations are interesting since they do not depend on the quality of the translation tools: translation tools - when they exist - are not efficient for all types of questions or fail to translate some named entities or multiword expressions. The paper is organized according to the following plan: first we describe the architecture of MUSCLEF (section 2), then we underline the difficulties encountered when shifting from a mono to a cross-lingual system (3), next we focus on the evaluation of this system and give the results obtained by three particular modules of MUSCLEF (4). We also give the general results of our participation to CLEF (5). Lastly, we present how we re-used different modules of MUSCLEF to build a first system for the Answer Validation Exercise (6).

2 System Overview

QALC, our monolingual system, is composed of four modules described below, the first three of them being classical modules of question answering systems: (a) the first module analyzes the question and detects characteristics that will enable us to finally get the answer: the expected answer type, the focus, the main verb and some syntactic features; (b) the second module is the processing of the collection of documents: a search engine, named MG¹, is applied; then the returned documents are reindexed according to the presence of the question terms, and finally a module recognizes the named entities and each sentence is weighted according to the information extracted from the question; (c) the third module is the answer extraction which applies two different strategies depending on whether the expected answer is a named entity or not; (d) the fourth module is the fusion of two sets of answers. Indeed our system QALC is applied on the Web as well as on the closed collection of the CLEF evaluation, then a comparison of both sets of answers is done; this way, we increase the score of answers that are present in both sets.

To build MUSCLEF, our cross-lingual question answering system, we added several modules to QALC, corresponding to both possible strategies to deal with cross-lingualism: question translation and term-by-term translation. In MUSCLEF, the first strategy uses Reverso² to translate the questions then our monolingual system QALC is applied. The second strategy, that we named MUSQAT, uses different dictionaries to translate the selected terms (a description and an evaluation of this translation are given section 3). Finally, we apply the fusion module to the different sets of answers: a first one corresponds to MUSQAT, a second one corresponds to the application of QALC on the translated questions, both these sets of answers coming from the CLEF collection of documents, and a third one corresponds to the application of QALC on the translated questions using the Web. MUSCLEF is presented Figure 1, where the first line of modules corresponds to our monolingual system QALC and the second line contains the modules necessary to deal with cross-lingualism.

3 Shifting from a Monolingual to a Cross-Lingual System

3.1 Performance Comparison

After the CLEF 2005 evaluation, CLEF organizers gave the original set of questions written in *good English* to the participants. Thanks to this new set of questions we could compare the behaviour of our different implementations: monolingual QALC, cross-lingual QALC (using Reverso), and cross-lingual MUSQAT. The results are given in Table 1. The results of document selection and document processing were calculated for 180 questions instead of 200 because of the 20 NIL

¹ MG for Managing Gigabytes, <http://www.cs.mu.oz.au/mg/>

² <http://www.reverso.net>

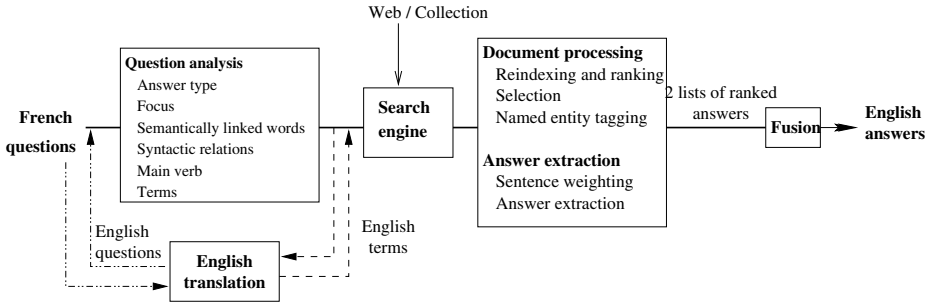


Fig. 1. Architecture of MUSCLEF, our cross-language question answering system

Table 1. Comparison of monolingual QALC, cross-lingual QALC and MUSQAT

	Monolingual system QALC	Cross-lingual systems QALC + Reverso MUSQAT	
Document selection	94.4	88.3	84.4
Document processing	93.3	87.7	82.2
Five sentences	67.5	58.5	50.5
Five short answers	40	39.5	36.5
First short answer	28	26	23

questions. Each number in this table represents the percentage of questions for which a good document/sentence/answer is returned.

Concerning the first three lines, we observe a big difference between the monolingual and the cross-lingual systems (from 6 to 17 %). This difference is due to missing translations: for instance acronyms or proper names (whose original alphabet can be different from ours) are often not correctly translated. In the last two lines, the differences are more surprising (and are due to problems in the answer extraction module): the monolingual system lost 40% of good answers during answer extraction, while the best cross-lingual system, QALC+Reverso lost 32.5%, and MUSQAT lost 27.7%. The results of the monolingual and cross-lingual systems on short answers are thus quite close. On the same CLEF 2005 data, [5] made also this kind of comparison: they report a loss of 24.5% of good answers between their monolingual French system QRISTAL (which obtains very high results: 64%) and their English-to-French system.

3.2 Corpus-Based Translation Validation

In this section, we present how in MUSQAT proceeds for the term and multiterm translation and their validation. The translation is done by using two dictionaries, Magic-Dic and FreeDict³, both being under GPL licence. Thus, the system MUSQAT gets several translations for each French word, which can

³ <http://magic-dic.homeunix.net/> and <http://freedict.org/en/>

be either synonyms or different translations when the term is polysemic. The evaluation made last year (reported in [6]) on term translation encouraged us to enhance our approach by the validation of these translations. To proceed to this validation, we used Fastr⁴ and searched a subset of documents (from 700 to 1000 documents per question) of the CLEF collection for either the bi-terms or syntactic variants of them. When neither a bi-term translation nor a variant was found, we discarded the corresponding translated terms. For example, to the French bi-term *cancer du sein* corresponded the three following translations: *breast cancer*, *chest cancer* and *bosom cancer*. In the documents only the first translation is present, which leads us to discard the terms *chest*, *bosom* and their corresponding bi-term. We hoped this could decrease the noise due to the presence of wrong translations. Unfortunately, this first experience in translation validation was not convincing, since we obtained nearly the same results with or without it. (22% of good answers without the validation, 23.5% with it). Undoubtedly this approach needs to be enhanced but also evaluated on larger corpora. Indeed, we only evaluated it on the corpus of CLEF 2005 questions, on which we obtained the following figures: from the 199⁵ questions, we extracted 998 bi-terms from 167 questions and 1657 non empty mono-terms; only 121 bi-terms were retrieved in documents, which invalidated 121 mono-terms and reduced the number of questions with at least one bi-term to 98. The number of invalidated mono-terms (121) is certainly not high enough in this first experiment to enable MUSQAT to reduce the noise due to wrong translations.

3.3 On-Line Term Translation

Yet, after the translation and its validation, some terms and multiterms remain untranslated. Web resources were then used to help find the missing translations, such as on-line dictionaries like Mediaco, Ultralingua or other dictionaries from Lexilogos⁶, the free encyclopedia Wikipedia, and the web site for European languages and cultures Eurocosm. Many of the terms that remain untranslated are multiword terms requiring a special strategy which is composed of three steps. First, all the multiword terms are cut into single words. Then the Web is browsed to get pages from all the on-line resources that contain these words. Each page is mapped into a common format which gives for each term its different translations. Finally, for each term of the original list, all exact matches are searched in the Web pages, and the most frequent translation is chosen. Table 2 shows an example of a mapping for the French term "voiture" and Table 3 the frequency of each of its translations. To avoid incorrect translations, only the translations of the exact term are considered. For the term "voiture", the most frequent translation is "car" and thus this translation is chosen. The results are

⁴ Fastr (<http://www.limsi.fr/Individu/jacquemi/FASTR/>), which was developed by Christian Jacquemin, is a transformational shallow parser for the recognition of term occurrences and variants.

⁵ 199 instead of 200 because one has been thrown out by the process.

⁶ www.lexilogos.com

Table 2. Translations of the French term *voiture*

French term	Translation
voiture	car
voiture	carriage
voiture d'enfant	baby-carriage
...	...
voiture	car
voiture	automobile
voiture de fonction	company car
...	...
voiture	car
...	...
clé de voiture	car key
voiture	automobile

Table 3. Frequency of the different translations of *voiture*

Translation	# of occurrences
car	3
automobile	2
coach	1
carriage	1

Table 4. On-line term translation results

Corpus	CLEF 2005	CLEF 2006
# of translated terms	195 (33%)	408 (32%)
# of untranslated terms	394 (67%)	858 (68%)
Total # of terms	589	1266

summed up in Table 4: for each corpus, about 30 % of the originally untranslated terms were translated thanks to this new module.

4 Evaluation of MUSCLEF Modules

4.1 Question Analysis

The question analysis module determines several characteristics of the question among which its category, expected answer type (named entity or not) and focus. We conducted a corpus study in order to validate our choice concerning these characteristics, and the focus in particular, on the corpus of English questions and collection. For the focus, we found that 54% of the correct answers contain the focus of the question, while only 32% of the incorrect answers do (against 20% and 11% for an non-empty word chosen by chance in the question), which tends to validate the choice we made for the focus. The performance of this module was evaluated in [7] which estimated its precision and recall at about 90% in monolingual tasks. The performance is lower on translated questions, since the question words or the structure of the question can be incorrectly translated. For example, the question *Quel montant Selten, Nash et Harsanyi ont-ils reçu pour le Prix Nobel d'Economie ?* (How much money did Selten, Nash and Harsanyi receive for the Nobel Prize for Economics?) is translated into *What going up Selten, Nash and Harsanyi did they receive for the Nobel prize of economy?*, which prevents us from determining the right expected answer type: *FINANCIAL_AMOUNT*.

4.2 Answer Extraction

In MUSQAT and QALC, the same method was used to extract the final short answer from the candidate sentence. In both systems, this last step of the question

answering process entails an important loss of performance. Indeed, in MUSQAT and QALC the percentage of questions for which a candidate sentence containing the correct answer is ranked first is around 35%, and as seen in section 3 the percentage of questions for which a correct short answer is ranked first falls to around 25%. During this step, about one third of good answers is lost. In [7], we exposed the reasons of the low performances of our answer extraction module. The patterns used to extract the answer when the expected type is not a named entity have been improved for the definition questions. In our last test, indeed, 21 questions among the 48 definition questions of CLEF 2005 were correctly tagged by the patterns. But in other cases, patterns still show a very low efficiency, for here linguistic variations are more important and remain usually difficult to manage.

4.3 Fusion

Since we now have three sets of results to merge, we proceeded in two steps: we first merged the results of QALC+Reverso and MUSQAT, which gave us our first run. And, as a second run, we merged our first run and the set obtained with QALC+web system. On CLEF 2005 data, the fusion step was not really convincing in terms of results: while QALC+Reverso gave 26 % of good answers and MUSQAT 22.5 %, Run1 gave only 25 % and Run2 27 %. Nevertheless, as we can see in section 5, on CLEF 2006 results the second fusion using the web gave better results since we obtained 25 % of good answers with the web and 22 % without. Concerning the first fusion, both systems (QALC+Reverso and MUSQAT) giving similar results, it is not surprising that the fusion does not increase the number of good answers. However, our fusion algorithm (described in [8]) is mainly based on the scores attributed by the different systems to their answers, and does not take into account the performances of the systems themselves, which could be a interesting way to improve it.

5 Results

Table 5 reports the results we obtained at the CLEF 2006 evaluation. As described just above (sub-section 4.3), we remind that the first run is the result of the fusion of two systems: QALC+Reverso and MUSQAT, while the second run is the result of the fusion of this first run and of QALC+Web. Last year, the best of our runs obtained a score of 19%, so the improvements brought to our

Table 5. MUSCLEF results at CLEF 2006

	Run 1	Run 2
First short answer	22.63%	25.26%
Confidence Weighted Score (CWS)	0.08556	0.15447
Mean Reciprocal Rank Score (MRR)	0.2263	0.2526

systems can be considered as encouraging. The difference of results between both runs strengthens the idea that the use of an external source of knowledge is an interesting track to follow. We underline, that at the time of writing this paper, only the first answer of each question has been assessed, so the MRR score does not bring more information than the number of good first short answers. The four first lines of results concern 190 questions, the 10 remaining questions were list questions for which the score is on the last line.

6 Answer Validation

In order to build the Answer Validation system, we used our QA system, applied to the hypotheses and justifications rather than to the questions and the collection, and we added a decision module. Our goal was to obtain the information needed to decide whether the answer was entailed by the text proposed to validate it. First the initial corpus file is processed to obtain the input format of our system, then the QA system is used to extract needed information from it, like a tagged hypothesis, a tagged justification snippet or terms extracted from the question. They are written in an xml file passed to the decision algorithm. We also get the answer our QA system would have extracted from the proposed justification, which is used to see if the answer to judge is likely to be true. Then, the decision algorithm proceeds in two main steps. During the first one, we try to detect quite evident mistakes, such as the answers which are completely enclosed in the question, or which are not part of the justification. The second step proceeds to more sophisticated verifications: (a) verifying the adequate type of the expected named entity if there is one; (b) looking the justification for terms judged as important during the question analysis; (c) confirming the decision with an extern-justification module using the latest version of Lucene to execute a number of coupled queries on the collection, like proximity queries (checks if a number of terms can be found close to one another within a text); the top results of each couples queries are compared in order to decide whether the answer is likely to be true or not; (d) comparing the results that our answer-extraction module (part of our QA system) would provide from the justification text. The results obtained by these different verifications are combined to decide if the answer is justified or not and to give a confidence score to this decision. Some errors have been corrected after submitting our results to the AVE campaign (which were rather bad, with very few positive answers). The results are given in Table 6 for the 3064 pairs judged during the evaluation campaign (201 pairs received an *unknown* evaluation). The “YES” (resp. “NO”) column corresponds to the pairs which our system judged as justified (resp. not justified). Among the *Correct NO*, those obtained during the first step previously presented, for which our system was sure of the answer (and judged as “NO” with the maximum confidence score 1), were distinguished from the others. The same distinction was established between the *Incorrect NO*. Among the 138 Incorrect “sure” NO, we observed that 63 were *badly* judged validations: the answer is correct, but the text given to justify this answer does not give any validation. For example,

Table 6. AVE results at CLEF 2006

	YES	NO		
Correct	177	2291	sure not sure	1370 921
Incorrect	68	528	sure not sure	138 390
Precision	0.72	0.81		
Recall	0.25	0.97		

the pair with id= "5496" is considered as validated, while the document given as a validation *des agresseurs de Naguib Mahfouz* does not contain the answer *EGYPTE* (which is nevertheless present in the hypothesis *Naguib Mahfouz a été poignardé à ÉGYPTE*).

7 Conclusion

Our cross-lingual system MUSCLEF presents the particularity to use three strategies in parallel: question translation, term-by-term translation and the use of another source of knowledge (actually limited to the Web). The three sets of answers are finally merged thanks to a fusion algorithm proceeding on two set of answers at the same time. The term-by-term strategy gives lower results than the most widely used strategy consisting in translating the question into the target source then applying a monolingual strategy. Nevertheless, we think it remains interesting from the multilingualism point of view, and we try to improve it by using different techniques of translation (use of several dictionaries and on-line resources) and validation.

References

1. Laurent, D., Séguéla, P., Nègre, S.: Cross lingual question answering using qristal for clef 2006. In: CLEF Working Notes, Alicante, Spain (2006)
2. Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., Vidal, D.: Priberam's question answering system in a cross-language environment. In: CLEF Working Notes, Alicante, Spain (2006)
3. Sutcliffe, R.F.E., White, K., Slattery, D., Gabbay, I., Mulcahy, M.: Cross-language french-english question answering using the dlt system at clef 2006. In: CLEF Working Notes, Alicante, Spain (2006)
4. Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedemann, J.: The university of groningen at qa@clef 2006: Using syntactic knowledge for qa. In: CLEF Working Notes, Alicante, Spain (2006)
5. Laurent, D., Séguéla, P., Nègre, S.: Cross lingual question answering using qristal for clef 2005. In: CLEF Working Notes, Vienna, Austria (2005)
6. Ligozat, A.-L., Grau, B., Robba, I., Vilnat, A.: Evaluation and improvement of cross-lingual question answering strategies. In: Workshop on Multilingual Question Answering, EACL, Trento, Italy (2006)

7. Ligozat, A.-L., Grau, B., Robba, I., Vilnat, A.: L'extraction des réponses dans un système de question-réponse. In: TALN Conference, Leuven, Belgium (2006)
8. Berthelin, J.-B., Chalendar, G., Elkateb-Gara, F., Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I., Vilnat, A.: Getting reliable answers by exploiting results from several sources of information. In: CoLogNET-ElsNET Symposium, Question and Answers: Theoretical and Applied Perspectives, Amsterdam, Holland (2003)

MIRACLE Experiments in QA@CLEF 2006 in Spanish: Main Task, Real-Time QA and Exploratory QA Using Wikipedia (WiQA)*

César de Pablo-Sánchez¹, Ana González-Ledesma²,
Antonio Moreno-Sandoval², and Maria Teresa Vicente-Díez¹

¹ Departamento de Informática
Universidad Carlos III de Madrid
{cesar.pablo,teresa.vicente}@uc3m.es

² Laboratorio de Lingüística Informática
Universidad Autónoma de Madrid
{ana,sandoval}@maria.111f.uam.es

Abstract. We describe the participation of MIRACLE group in the QA track at CLEF. We participated in three subtasks and presented two systems that works in Spanish. The first system is a traditional QA system and was evaluated in the main task and the Real-Time QA pilot. The system features improved Named Entity recognition and shallow linguistic analysis and achieves moderate performance. In contrast, results obtained in RT-QA shows that this approach is promising to provide answers in constrained time. The second system focus in the WiQA pilot task, that aims at retrieving important snippets to complete a Wikipedia. The system uses collection link structure, cosine similarity and Named Entities to retrieve new and important snippets. Although the experiments have not been exhaustive it seems that the performance depends on the type of concept.

1 Introduction

In our third participation in CLEF@QA task, the MIRACLE group took part in the main task and two pilot task, Real Time-QA and WiQA. All the submissions used Spanish as source and target language. We submitted two different runs for each of the tasks.

The QA system and the core set of tools have been developed along previous evaluations[3]. We have enhanced all the modules in terms of functionality and architecture. As of functionality, we have improve the recognition, classification and normalization of complex expressions like Named Entities and Temporal Expresions. We have explored different ways to take into account them in the main QA task and their role to signal important and new information in the WiQA task.

* This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and two projects by the Spanish Ministry of Education and Science (TIN2004/07083 and TIN2004-07588-C03-02.)

The system has also been adapted to some extent to the new requirements of 2006 QA task[5]. Questions no longer are provided with a question type and therefore the system should infer it. The normalization of timexes was partially accomplished for the evaluation submission and it also improved results for temporally restricted questions. The guidelines also required short passages supporting the answers. New questions types were also introduced in the main task, in particular list of factoids. Although our system is prepared to categorize all the new types, the extraction for the List type worked as in factoids which have shown insufficient. The same system was used for the Real Time QA task that consider not only accuracy but also takes into account the time to produce and answer.

As a result of the architectural improvements, the system is modular enough to make easier the reuse of different modules. A basic system for the WiQA task[13] was assembled in less than two weeks. In a few words, the aim of the task is completing a Wikipedia article about a topic with new and important information from other pages. The task has some resemblance to other information access tasks like 'Other' questions in TREC evaluations[4], novelty detection[7] and multi-document summarization.

The task uses the WikipediaXML [8] collection which is an XML version of Wikipedia produced for several languages. It is also a semistructured document collection with rich link structure. It includes links between articles in the same language subcollection and also links between articles in different languages. The systems should locate new information that is relevant enough for the topic to be included in the Wikipedia article.

2 System Description

Our toolkit is composed of a set of language analysis tools that are used in several applications and task specific modules. The same QA system was used in the main task and Real-Time QA pilot task with exactly the same configuration. The system is organized in the classic pipelined architecture and performs Question Analysis, Information Retrieval and Answer Extraction.

A different system was built for the WiQA task that only performs an specific kind of information retrieval task, that is subdivided in Candidate Selection, Novelty Detection and Ranking.

2.1 Language Analysis

The three modules of the QA system and the WiQA Novelty Detection system uses the same tools for Language Analysis.

The main component of the Language Analysis module are DAEDALUS STILUS[12] analyzer service. STILUS provides tokenization, sentence detection and token analysis for Spanish. The analysis of tokens include information about POS tags, lemmas and other morphological features such as number and tense. The analysis is also enriched with semantic information which is stored in a dictionary of common Named Entities (NE) extracted from several resources. Those

NE are organized following Sekine's typology[11]. Thematic relations and geographical relations for some of those NE are also given although we have not used them in the QA system yet. This tool has been developed for spell, grammar and style checking in mind and therefore is very exhaustive in language coverage and their analysis, even if some are infrequent uses. As this tool is reused for this task it does not include a proper POS tagger or chunking like other text analysis toolkits focused in Information Extraction tasks. In contrast it has excellent performance for large quantities of text and is fast, stable and robust.

In order to adapt the output of STILUS to the processing requirements of the QA task we have added some more language analysis modules. A very basic set of rules remove infrequent uses of some common words based on contextual information. To improve recognition of NE, in particular recall, another set of rules are used. This step recognizes and groups tokens based on the previous information and other orthographic, contextual and structural features. Most semantic information is retained and even some rules could add new entity class information to the analysis. This module has been revised and improved since last year submission. We have also include some improvements in Temporal Expressions (TE) recognition and normalization that for the time of our submissions was in a preliminary stage but seem to be effective for common cases.

2.2 QA Architecture

The same system was used in the QA main task and Real-Time QA. Linguistic analysis can be performed on-line or off-line. On-line processing is performed only on the candidate documents, while off-line processing is applied to the whole collection. We have used full on-line processing for all our submissions. The following sections explain in more detail each of the modules.

Question analysis. This module has been extended to deal with new kind of questions and the fact that the question type would not be given as an explicit input to the systems.

The question analysis module transform the question string into a common representation used throughout the system. Our question model is very simple as we lack complex analysis resources. We characterize each question with the following features[9]: question type (*QT*), expected answer type (*EAT*), question focus (*QF*), answer type term, query terms and relevant terms. Query terms (*Qterms*) are those considered to retrieve candidate documents and are used to build queries. On the other hand, relevant terms (*Rterms*) are a broader set of terms that could help to locate an answer but would retrieve a lot of noisy documents if used in the queries.

Question classification is carried in two steps. First, question is classified regarding their QT and later on the EAT is assigned. We have used a set of handwritten rules to perform question type classification. Four question types are detected, Factoids (F), Temporally Restricted Factoids (T), Lists (L) and Definitions (D). Besides common patterns, the classification use specific information for any of the types. For instance, the detection of TE usually signals for T questions and the use of plurals usually signals L questions.

The EAT is also assigned using rules and lists of related words generated by a linguist. This year we improved coverage for this classification by adding more patterns and extending the lists. F,L and T questions share the rules and the same hierarchy of EAT used last year[3]. For D questions the type reflects the object to describe or action to take (Person, Organization, Other or Acronym expansion) and uses mainly the information of our NE recognition module.

Document retrieval. We used Xapian[14] for the retrieval stage. Xapian is an open source engine based on the probabilistic retrieval model that includes the Okapi BM25 scoring function [6]. The collection has been indexed using simple terms stemmed with Snowball stemmer[10] for Spanish and query terms are processed in a similar fashion.

Simple *Qterms* are used in the query and composed with boolean operators. First N_d candidate documents are selected by the retrieval step and analyzed. All the sentences that contain a number of relevant terms from the questions are considered for further processing. We have experimented with different criteria to select sentences based on how to account for Named Entities and other complex terms.

Answer selection. We have developed filters for every EAT in our system that identify candidate answers from the set of relevant sentences previously identified. A local score is calculated for every pair of answer and sentence considering the document score and the frequency of *Rterms*. Every term receives credit in inverse proportion to their frequency in the set of selected sentences for a question. We consider two methods regarding complex terms, the first computes the score using complex terms without decomposing while the second decompose them and use the complex terms and all the simple ones except stopwords.

In a second step, similar candidate answers are conflated and a new global score is generated. When grouping several candidate answers, the one with the higher score is selected as representative for that group, together with its snippet. A redundancy score is calculated by counting the number of snippets in a group up to a maximum of N_r . A global score is assigned to the group that combines the local score (using weight *wls*) and the normalized redundancy score (*wrs*) to produce a confidence between 0 and 1. The values used in our experiment has been assigned manually and are $N_r = 10$, $wls = 0.9$, $wrs = 0.1$.

2.3 WiQA Architecture

For solving the WiQA task, we consider that a system should have at least these three modules: Candidate selection, Novelty detection and Importance Ranking. The evaluation and retrieval unit for the task is the sentence, and therefore most of the operations proceed at this level. As the system was developed in a short period of time, most of the effort was given to the novelty detection step.

Candidate Selection. This module selects sentences that are related to the article that we are completing. The article is described by a title, which for an specific language collection is unique as the Wikipedia uses the title as the id. There are several alternatives to retrieved reated content but in our system we rely on the link structure between articles, in particular *inlinks*. We call inlinks those internal links (which use tag *collectionlink*) that refer to the query article. This approach is very precise and selects sentences that are clearly related to the query because the link is unambiguous. On the other hand, not all mentions are linked to an article, so this approach could leave important information out.

The collection has been stored using an open source XML database, Berkeley DBXML [1]. We use XQuery to retrieve relevant passages that have inlinks to the query article. Several structural indexes have been created to support efficient queries. We have experimented with diferent sizes for the passages, using sentences and paragraphs marked in the XML structure.

Novelty Detection. This module filters sentences that are near duplicates to those already in the article. We have implemented novel sentence detection by a combination of two measures, cosine similarity between sentences and the ratio of novel an already present Named Entities (NE).

Cosine similarity has been proposed as a simple and effective measure to detect similarity between sentences [2]. To calculate cosine similarity we index the original article in a sentence by sentence basis and compare each of the candidate sentences to them.

Our second novelty measure is based on the presence and absence of NE in a broad sense, considering quantities and temporal expressions. We believe that the presence of these facts is probably a signal for important information in reference texts like Wikipedia. For every candidate sentence, we obtain three related measures, the number of NE that appear in the article, the number of new NE and the ratio.

The final decision step uses manually defined thresholds that were obtained in a small and separate development set. The first threshold requires a minimun number of common NE to consider the sentence. Another threshold filters sentences with high similarity. Finally, the third threshold filter those sentences in which the ratio of new NE is not significant.

2.4 Ranking

The third module should present the sentences in a appropriate order to the user. We have used the information that we had available at that point and we have tried different combinations of the two previous measures, NE ratio of novel entities and cosine similarity. These measures have been obtained using a small development set, and again a more principled way of scoring should be considered.

3 Results

3.1 QA Main Task

Two runs were submitted to the Spanish monolingual task, they differ in the way they weight multiwords or complex terms like NE, TE and numerical expressions. As we index the EFE collection using simple terms it was not clear which strategy works best for selecting and scoring sentences. We have tried two different simple alternatives in each of the runs. Run mira01eses selects and scores sentences considering the whole multiword which favour high precision. In contrast, run mira02eses decomposes the term and use in addition simple terms which favour recall.

Table 1 presents the official results for factual (F), definition (D) and temporally restricted questions (T) for the two monolingual spanish runs submitted.

Table 1. Evaluation results for MIRACLE main task runs

Run	R	X	U	Acc(%)	F(%)	T(%)	D(%)	L(%)	NIL-F	r
mira061eses	37	3	6	18.50	21.30	15.00	16.67	10.00	0.34	0.136
mira062eses	41	4	7	20.50	21.30	17.50	23.81	10.00	0.35	0.145

Performance for the two runs is very similar although the runs differ moderately between them. Run mira062, which decomposed complex terms and use the term and its decomposition, obtain better score specially in T and D questions. This fact should be taken with care as just the first answer has been evaluated and we believe that more significant differences would appear we consider a larger number of answers per question. Compared with our last year performance, we have improve the accuracy for T questions. In contrast, D questions are of increasing difficulty and are responsible for the loss in performance together with L questions for which we did not designed an special strategy.

The correlation between the confidence score given by the system and the correctness it is also better for the run mira062 but much lower than last year. Problems with the confidence function remains as it often assign the same score to several candidates which results in random selection and a loss in accuracy.

3.2 Error Analysis

It is important to estimate the causes for low scores got this year, analysing which module in the general system is the main responsible for the errors. Of course errors in a QA system cannot be assigned to only one subsystem, and this should be beared in mind when judging the following error analysis. The estimation of errors for run mira061eses are shown in the Table 2. In the first column, we show the percentage of error types with respect to the total of errors considering Wrong, ineXact and Unsupported (in the case of mira061eses, 154 questions). It seems that most of the errors these year are dued to the document retrieval stage.

Table 2. Error analysis for run mira061

Error type	% errors
Question classification	7,79
Question analysis	18,18
Document retrieval	38,96
Sentence selection	3,89
Answer extraction	11,03
Answer ranking	20,12

3.3 Real-Time QA

It is acknowledged that some of the interesting scenarios for a QA system impose time constraints that limit the processing and resources a system can use to answer a question. This pilot task which was performed live during the CLEF 2006 workshop explores the trade off between answer accuracy and processing time. The systems answered 20 question in monolingual Spanish and the time taken was recorded.

We submitted two runs with the same configuration than mira062eses, using on-line document processing. In the current configuration our system analyzes the whole document and therefore it was natural to use the number of documents retrieved N_d as the independent variable. The first run used 100 documents while the second run used only 40 documents. After the workshop we have performed additional experiments using only 10 documents and have measured the times more accurately in the same question set. We have also changed the system to use the EFE collection preprocessed with STILUS. NE detection and filtering is still performed on-line. The system runs in a mixture of Java and C++ (Xapian, STILUS) in a Linux laptop with a 1,7GHz and 1GB of memory.

Table 3. Evaluation results for MIRACLE Real Time QA runs

			Time(s)	
Run	N_d	MRR	on-line	off-line
official run 1	100	0.41	549(496)	156
official run 2	40	0.33	198(182)	73
unofficial run 3	10	0.30	48	14

The results show that Real-Time QA could be achieved with a simple architecture like the one presented. Although not realistic for moderate size and large collections, in the case of corpus preprocessing the system could answer in less than a second per question if using only the first 10 documents. The time for processing a questions is reduced in a factor of three or four. Besides the need for obvious improvements like processing in a sentence by sentence basis, the results support the idea that research on specific approaches of IR for QA is needed to maintain accuracy scores in RT-QA.

3.4 QA over Wikipedia

Our group submitted two runs to the WiQA subtask using the Spanish subcollection. Both runs used the same modules but in a different configuration. The main difference between runs are the length of the context that is considered for candidate selection. The first run mira-IS-CN-N uses only sentences that link to the article while and the second run mira-IP-CN-CN considers the whole paragraph related. Parameter and ranking function were tuned manually for the two runs.

The Spanish subtask consisted in a total of 67 topics that were developed by participant teams. The 50 first topics were correctly tagged as PERSON, LOCATION or ORGANIZATION and had a distribution of length from stubs to articles. The other 17 topics were considered optional but as category information was not considered results are similar. We report results for the official topics in Table 4. The average yield of sentences is about one for each of the methods being run mira-IP-CN-CN more precise as it returns less results. In general, both methods return only a small number of sentences. We believe that this is the effect of the candidate selection method, that uses only links, and the structure of the links in the Spanish collections. A method using content retrieval would have probably returned more candidate sentences.

Table 4. Evaluation results

<i>Run</i>	<i>#topics</i>	<i>#snippets</i>	<i>#i</i>	<i>#i + n</i>	<i>#i + n + nr.</i>	<i>Av. Yield</i>	<i>MRR</i>	<i>P@10</i>
mira-IS-CN-N	50	176	96	62	54	1.08	0.24	0.31
mira-IP-CN-CN	50	310	115	74	52	1.04	0.24	0.17

In Table 5, we analyze the performance of the runs regarding the different class of topics from the official set. Topics where classified as Person (P), Location (L) and Organization (O). Significant differences are appreciated in each of the classes for the two runs. For P topics the run based on paragraphs performs much better, while the results are completely different for L topics. We believe that the difference arise as paragraphs that link to a location are not always related and locations often appear in tables and lists.

Table 5. Results by NE category for the official topics

<i>Run</i>	<i>#category</i>	<i>#topics</i>	<i>Av. Yield</i>	<i>MRR</i>	<i>P@10</i>
mira-IS-CN-N	P	17	0.76	0.17	0.30
mira-IP-CN-CN	P	17	1.47	0.31	0.22
mira-IS-CN-N	L	17	1.12	0.25	0.29
mira-IP-CN-CN	L	17	0.47	0.15	0.08
mira-IS-CN-N	O	16	1.38	0.31	0.33
mira-IP-CN-CN	O	16	1.19	0.27	0.20

4 Conclusion and Future Work

We have described our QA system developed for the main QA task and the Real-Time QA pilot. The system has featured several improvements in all of their modules but there has not been a general remarkable improvement in their performance compared with last year. In contrast, for some question types and some answer types we have obtained singular improvements.

Most of the errors have been identified as part of the Retrieval and Answer selection modules. Visual inspection of the results and feedback from casual real users have shown that extractions are still very noisy and a source of distrust. Besides being brittle, answer extraction is also one of the most time consuming task for developing a QA system. Further work is expected to approach these problem by learning patterns for extraction and filtering candidates. A second problem that remains unsolved is ranking and confidence estimation in a principled way. On the other hand, retrieval also need to be addressed and we planned to experiment with passage in addition to sentence selection. All these changes and other performance improvements are in keeping with the objective of providing answers in real time.

We also presented a second system that uses the same core elements to solve a different but related task, WiQA. We believe this task could help to select important information about a topic even if the user does not know enough about it to formulate a question. This could be helpful in an interactive QA system. A problem that we already foresaw while building the system is related to the initial retrieval of candidate sentences, which in our case use link structure. During the workshop differences in the distribution of links for the Spanish, English and Dutch topics seemed to matter for these initial step. Nevertheless, the most important problem of this task consist on modelling importance for a topic and these is implicit to the difference performance for difference types we hae obtaines.

An interesting fact about this task, it is that given a semistructured corpus like WikipediaXML it seems easy to build a multilingual system and it provides and interesting scenario for cross-lingual retrieval.

References

1. Sleepycat berkeley db xml 2.2 (last visit July 2006), On line <http://www.sleepycat.com/products/bdbxml.html>
2. Metzler, D., Bernstein, Y., Croft, B., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 517–524. ACM Press, New York (2005)
3. de Pablo-Sanchez, C., et al.: Miracle's 2005 approach to cross-lingual question answering. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Voorhees, E.M., Dang, H.T.: Overview of the trec 2005 question answering track. In: Proceedings of the Fourteenth Text REtrieval Conference (2005)

5. Magnini, B., et al.: Overview of the clef 2006 multilingual question answering track. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain. LNCS, Springer, Heidelberg (2006)
6. Robertson, S.E., et al.: Okapi at trec-3. In: Harman, D.K. (ed.) Overview of the Third Text REtrieval Conference (TREC-3) (1995)
7. Soboroff, I.: Overview of the trec 2004 novelty track. In: The Thirteenth Text Retrieval Conference (TREC 2004) (2004)
8. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)
9. Pasca, M.: Open Domain Question Answering from Large Text Collections. CSLI Publications, Stanford (2003)
10. Porter, M.: Snowball stemmers and resources website (last visited, July 2006), On line <http://www.snowball.tartarus.org>
11. Sekine, S.: Sekine's extended named entity hierarchy (last visited August 2006), On line <http://nlp.cs.nyu.edu/ene/>
12. Stilus website (July 2006), On line <http://www.daedalus.es>
13. Jijkoun, V., de Rijke, M.: Overview of wiqa 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain. LNCS (September 2006)
14. Xapian: an open source probabilistic information retrieval library (last visited July 2006), On line <http://www.xapian.org>

A First Step to Address Biography Generation as an Iterative QA Task

Luís Sarmento

LIACC - Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n, 4200-46 Porto, Portugal
las@fe.up.pt

Abstract. The purpose of this article is two-fold: (i) to present RAPOSA, an open-domain automatic question answering system for portuguese that participated in QA track at CLEF 2006 for the first time, and (ii) to explain how RAPOSA is intended to be a key component of a larger information extraction framework that uses question-answering technology as the basis for automatic generation of biographies. We will make a first attempt to classify questions regarding their relevance for iterative biography generation and, according to such classification, we will then explain our motivation for participating in CLEF 2006. We will describe the architecture of RAPOSA, explaining the internal details of each of its composing modules. Next, we will present the results RAPOSA obtained on this year's edition of the QA track, and we will identify and comment on the main causes of error. Finally, we will present directions for future work.

1 Introduction

It is generally accepted that today's QA systems are reasonably good in finding simple answers to several types of questions. For example, for a question like "Who is X?", with X being a name of a person, most of today's QA systems will easily find at least one "good" answer, usually containing information about the nationality and job description of person X. However, it is highly arguable that such an answer is satisfactory in a realistic scenario: most users would probably expect a more complete biographic profile about person X. In a certain sense, a question such as "Who is X?" can be seen as a series of implied biographic questions. A more user-friendly QA system would provide the user with additional information about X, such as X's place and date of birth, X's place and date of death (if that is the case) or information about X relatives (husband / wife / children / parents...) [1]. Ideally, such QA system would be able to formulate other more specific questions depending on X's profile. For instance, if X was known to be an sportsman, answers to other related questions such as "In which team did X play (in 1993)?" would probably be relevant for the information need initially expressed through the question "Who is X?". Using such an iterative question answering procedure, a system would eventually build a complete biographic profile of X, even if just composed of factoids. Our future goal is to

develop an automatic biography generator system for Portuguese using this iterative question answering strategy. There are several challenges involved in this approach: the issue is not only finding the correct answer(s) to a given question but also being capable of automatic formulating the right biographic-relevant questions.

Our main motivation for participating in QA@CLEF for the first time is (i) to develop a deeper insight regarding biographic questions and (ii) to test our QA system - RAPOSA - in answering such questions. In this year's participation, therefore, we made a first attempt in trying to answer some of those questions with RAPOSA, namely those regarding people (simple definitions), place, dates and quantities.

2 Biography-Related Questions

The participation in CLEF helped us to make an initial and informal study about the relevance of certain questions for the purposes of biography generation. We identified three different groups of questions regarding biographical elements.

The first group contains questions that refer to simple attributes common to all people, and that are generally relevant in *all* biographies. Such questions can immediately follow from a "Who is X?" question because they do not require any significant background knowledge or inference about the X person. We will say that these question belong to the *elementary question set*. Looking at the 200 questions set from 2006 QA@CLEF track, we were able to identify 13 question that can be considered part of such an elementary question set (excluding the "Who is X?" questions). The next list shows some examples, along with the question number provided by the organization:

1. Who is Elisabeth The Second's father? (0076)
2. What's the name of George's H. W. Bush's wife? (0105)
3. Which city was Wolfgang Amadeus Mozart born in? (0118)
4. In what year did Charles De Gaulle die? (0179)
5. Where did Kurt Cobain die? (0455)
6. Where does José Rodrigues Filho work? (0760)

There are other question regarding features that, although attributable to all people, may not be equally relevant for all biographies. For example, questions like "How tall is X?", or "What is X weight?", although applicable to any person, are definitely more relevant for a biography of a sportsman than they are for a biography of politician or a writer. On the other hand, from a biographical point of view it seems more natural to ask "Which party does X belong to?" if X is known to be a politician rather than if X is a sportsman, despite the fact that any person can have political affiliations. Such questions are only relevant for the biography of people belonging to certain profile types (e.g. artists, politicians, sportsmen, businessmen) so, for our purposes, we will name these questions as *profile dependent questions*. In 2006, the QA@CLEF question set has several question that could be considered profile dependent questions, such as:

1. When did the official coronation of Elizabeth The Second take place? (0086)
2. What motion picture studio was Cedric Gibbons chief art director of? (0356)
3. Who was the Prime Minister of England before John Major? (0416)
4. Which party does Mahfoudh Nahnah belong to? (0469)
5. Name a film directed by Claude Lanzmann. (0481)
6. What team does Johann Cruyff coach? (0756)
7. What does Vital do Rego teach? (0779)

As it can be immediately seen, the formulation of profile dependent questions involves a certain amount of background knowledge, and some inference over basic information about a person (e.g.: a job description). Such questions may focus, for example, on events or organizations related to the person's profile, or even to other people sharing a similar profile. For example, question 0416 from the previous list is important for the biography of John Major, since we might want to know who preceded him in is job, a fact that is quite relevant when the job is being a Prime-Minister.

There are still other questions that may involve some fruitful *speculation* regarding a given profile. It might be interesting for the QA system to assume some hypotheses and then generate questions based on those (unconfirmed) assumptions. For example, asking "When did X commit suicide?" can be formulated even if it is not confirmed that X did in fact commit suicide. However, if an appropriate answer is found to this question, two very important pieces of information are gathered at the same time. The CLEF 2006 question set has some examples such as "What year was Martin Luther King murdered?" (0114) or "When did Cleopatra commit suicide?" (0799). In some cases, if we take into account a known profile, these assumptions can be even more speculative, but still reasonable. For example, the following questions could be formulated as part of an exploratory strategy for finding information, based on reasonable assumption about the profile type: "Which symphony was composed by Beethoven in 1824?" (0063), "Where did Braque and Picasso work together?" (0348), "How many countries did Nixon visit from 1953 to 1959?" (0390) or "Where did Abba win the Eurovision Song Contest?" (0733). For our purposes, we consider these questions as belonging to the *speculative question set*.

3 RAPOSA

RAPOSA (Respondedor Automático a Perguntas que é Ouro Sobre Azul) is an open domain question answering system for Portuguese. Differently to other question answering systems for Portuguese that make use of extensive linguistic resources [2] or deep parsing techniques [3], RAPOSA uses shallow parsing techniques over the semantic annotation produced by a named entity recognition (NER) system, SIEMÊS [4]. Generally speaking, RAPOSA assumes that the correct answer for several types of question is one entity tagged by SIEMÊS, and its job is to select the right one. The importance of named-entity recognition in QA systems is widely accepted. According to [5], 88.5% of the Portuguese questions in QA@CLEF 2004 refer explicitly to named entities, while 60.5% of the answers

are either named entities or they include named entities. In fact, using a named-entity recognition system as a the major source of semantic information during answering extraction is a standard strategy in question answering systems (as evidenced in [6], [7], [8] or [9]).

RAPOSA is currently a pipeline consisting of 6 modules, and is the result of improvements made to the initial architecture presented in [10]. Each module will be described in turn.

3.1 Question Parser

The Question Parser operates in two steps. First, it invokes SIEMÊS in order to identify named entities in the question: people, places, organizations, dates, titles (e.g.: book titles). These are usually either the arguments of the question to be parsed or some important restrictions. SIEMÊS also helps to find other important elements, such as job descriptions, that may be relevant for parsing the question. In a second step, the Question Parser uses a set of 25 rules to analyze the type of question and to identify its elements. After the type of question has been found, these rules try to identify the arguments of the question, argument modifiers, temporal restrictions and other relevant keywords that may represent good clues for retrieving snippets from the answer collection. Depending of the type of question, the Question Parser also generates a list of admissible answer types that are compatible with the tagging capabilities of SIEMÊS. The output of this module is an object containing the parsed question in a canonical form. The next list illustrates such output, including the tags produced by SIEMÊS:

- Question: Quem foi o último presidente da Rússia antes de 1990 ?
- Question Type: Definition
- Class: Job description
- Question head: “Quem foi o”
- Arguments: “presidente” < job description >; “Rússia” < org >
- Modifiers: “último” < time rest >
- Restrictions: “antes de 1990” < period >
- Answer Type: < human >

Currently, these rules only address some types of questions, namely those that refer to people (“Quem...?” / “Who...?”), to places (“Onde...?” / “Where...?”), to time (“Quando...?” / “When...?” or “Em que ano...?” / “In what year...?”) and to quantities (“Quanto...?” / “How many...?”).

3.2 Query Generator

The Query Generator module is responsible for preparing a set of query objects from the previously parsed question so that the following Snippet Searcher module can invoke the required search engines. The Query Generator selects which words must necessarily occur in target text snippets, and which words are optional. The Query Generator may optionally perform a “dumb” stemming procedure over all words, except the words the belong to the arguments, to produce a less restrictive query. This helps to increase the number of snippets that

match that query, and hopefully improve global recall of RAPOSA (we were not yet able to assess the impact of this option in the global performance of the system). Currently, our stemming procedure is very basic: it simply replaces the last 2 or 3 characters of words to be “stemmed” for a wild-card.

3.3 Snippet Searcher

The Snippet Searcher is the module responsible for obtaining text snippets from which an answer may be looked for, using the query objects given by the Query Generator. The Snippet Searcher is intended to be the interface to web search engines and other text repositories available. Currently, the Snippet Searcher is only using two text databases: the CLEF document collection and BACO [11], a snapshot of the Portuguese web in 2003 [12].

The Snippet Searcher receives a set of query objects and transforms them into lower level search expressions. When using the CLEF document collection or BACO, the Snippet Searcher generates the appropriate SQL statement, since both collections were indexed using the MySQL database engine (text was indexed at the sentence level). Snippets obtained from the CLEF collection and from BACO range from 1 to 3 sentences. Each snippet keeps reference to its source, in order to justify the answers with the document ID (or URL) from which the snippet was extracted. For the QA@CLEF track, only the CLEF document collection was used for obtaining one sentence snippets. After retrieving the text snippets, the Snippet Searcher invokes SIEMÊS to NER annotate them.

3.4 Answer Extractor

The Answer Extractor takes as input the question in canonical form and the list of NER annotated snippets given by the Snippet Searcher. It assumes that the answer to the question is one of the elements tagged by SIEMÊS, and since the type of admissible answers for the question at stake has already been determined by the Question Parser, the answer is usually already quite constrained to the set of *semantically compatible* named entities (or other tagged elements) found in the snippets.

The Answer Extractor has two possible strategies to find candidate answers. The first one is based on a set of *context evaluation rules*. For each tagged text snippet, RAPOSA tries to match certain contexts around the position of the argument (note that the argument of the question must be present in the text snippets). For example, for a question like “Who is X?” with X being the name of a person, the answer is expected to be a job description so that checking the existence of those elements around occurrences of X in certain contexts might lead to the candidate answer. In this case, the rule might check for patterns like “... < job description > X ...” or “... X, < job description > ...” with < job description > standing for the element in the text snippet tagged by SIEMÊS as a job description. RAPOSA has currently 25 of such rules for dealing with questions of the type “Who is < job description >?” and 6 rules to deal with questions of the type “Who is < person name >?”. For RAPOSA’s participation

in CLEF 2006 we were not able to develop similar rules to deal with other types of questions.

The second strategy available for answer extraction is much simpler: it extracts as possible answer any element tagged with a semantic category that is compatible with the expected semantic category for the question at stake. For example, for a question like “When was < EVENT > ?” the Answer Extractor will collect all elements tagged as < date > in the text snippets (which match the string < EVENT >) provided by the Snippet Searcher. Although this strategy is potentially very noisy, since more than one compatible element may exist in the snippet, one expects the correct answer to be one of the most frequent candidates extracted, provided that there is enough redundancy in the answer collection. We call this the *simple type checking strategy*.

The output of the Answer Extractor, no matter which strategy is used, is a list of answer objects containing the candidate answer and the text snippets from which the answer was extracted. If no answer is found in the text snippets given by the Snippet Searcher, RAPOSA produces a NIL answer for the question, assigning a low value of confidence to the answer (0.25) to acknowledge the fact that such result may be due to a lack of better analysis capabilities.

3.5 Answer Fusion

The role of the answer fusion module is to cluster lexically different but semantically equivalent (or overlapping) answers in to a single “answer group”. At this moment, this module is not yet developed: it simply outputs what it receives from the Answer Extractor.

3.6 Answer Selector

The job of the Answer Selector is (i) to choose one of the candidate answers produced by the Answer Fusion module, (ii) to select the best supporting text snippets and (iii) to assign a confidence value to that answer.

Currently, for deciding which of the candidates is the best answer, the Answer Selector calculates the number of supporting snippets for each candidate and chooses the one with the highest number of *different* supporting snippets. The assumption behind this strategy is that by promoting candidates that are supported by many *different* snippets, we eliminate the bias produced by the presence of duplicate documents in the search collections. However, we need to further study the impact of this particular option on the global performance of RAPOSA, comparing it with results obtained when directly using the absolute frequency of the snippets.

Choosing the “best” supporting snippets is quite straightforward when the Answer Extractor uses the context matching strategy because the snippets found must match very specific patterns, which almost always have explicit information for supporting the answer. However, if the candidate answers are obtained using the *simple type checking strategy*, the Answer Selector has no way, at the moment, of deciding if a given snippet (containing both the argument and the chosen candidate answer) really supports the answer. So, in both cases, the procedure

is simply to randomly choose up to 10 different supporting snippets associated with the answer.

For the participation in QA@CLEF 2006, RAPOSA used a very simple rule for assigning the confidence level to answer. The chosen answer - which has at least one supporting snippet - was given a confidence level of 1.0. This value obviously disregards important information such as the number of alternative candidates available and the corresponding number of supporting snippets. After the CLEF event, we made a small improvement in this method, which nevertheless still ignores many important factors. RAPOSA now calculates the confidence level $c(a_i)$ for a given the answer a_i chosen from the set of all candidate answers $A = a_1, a_2, \dots, a_n$ using Equation 1. This formula takes into account the number of supporting snippets found and, indirectly, the number of alternative answers.

$$c(a_i) = \frac{\# \text{ snippets}(a_i)}{\sum_{k=1}^n \# \text{ snippets}(a_k)} \quad (1)$$

4 RAPOSA at CLEF06

We submitted two runs with two different RAPOSA configurations. It is important to say that during our initial study of this problem we believed that the definition questions - “Who is ‘person name’ ?” and “Who is ‘job description’ ?” - were the most important ones because they would allow us to determine (or check) the profile type of a given person. Therefore, during the earlier stages of the development of RAPOSA, we were mainly considering these types of questions, and most of the question parsing and answer extraction rules developed for QA@CLEF 2006 had these questions in mind.

The first run submitted, R1, was configured to extract candidate answers using the *context matching rules* developed, i.e. using the most restrictive (and hopefully the highest precision) strategy of the Answer Extractor. However, these rules only covered the 34 “Who is ...?” questions, among the 200 questions that were proposed. All other questions were automatically given a NIL answer with a confidence value of 0, to indicate that RAPOSA did not even try to answer them. Since it was quite disappointing to see RAPOSA ignoring over 80% of the questions (we initially expected the number of “Who” questions would be higher), we made a second attempt to answer other biography-relevant questions, namely those regarding dates, locations and quantities. As it was not possible to develop *context matching rules* for all those cases, we decided to use the more relaxed *simple type checking strategy* to extract the answers for those questions.

Therefore in the second run, R2, RAPOSA was configured to use two alternative strategies to extract answers, chosen according to the type of the question at stake. Thus, for the 34 “Who is...?” questions, RAPOSA was configured to behave exactly as in R1, and perform answer extraction using the same *context matching rules*. For the other questions - place (“Onde...?” / “Where...?”), time (“Quando...?” / “When...?” or “Em que ano...?” / “In what year..?”) and quantities (“Quanto...?” / “How many...?”) questions - RAPOSA was configured to

extract answers using the *simple type checking strategy*. Again, all other questions would immediately receive a NIL answer with the corresponding value of confidence set to 0. In R2, RAPOSA was able to answer 40 more questions than in R1, making a total of 74 questions.

We have manually checked all the answers against the list of correct answers provided by the organization, considering an answer correct if and only if it exactly matches the answer provided by the organization. We make the distinction between three types of cases among the questions that RAPOSA tried to answer (their confidence values helped to identify the cases):

1. ANS: questions that were answered and supported.
2. NIL1: questions to which RAPOSA did not find any answer after analyzing snippets provided by the Snippet Searcher (NIL + confidence level = 0.25).
3. NIL2: questions to which RAPOSA was not able to find any snippet in the document collection to search for the answer (NIL + confidence level = 1.0).

The results of both runs are given in Table 1, along with the corresponding values of precision *per answer type* ($P_{type} = right_{type}/total_{type}$) and global precision (34 questions in R1, and 74 questions in R2). The last group of columns presents explicitly the performance of the type checking strategy used to answer place, time and quantity questions in run R2 (i.e. those questions not answered in R1).

The precision of R1 is not as high as it was initially expected: a global value 0.18% in answering “Who is ...?” questions seems disappointing. Looking in more detail at the 11 wrong answers produced by RAPOSA in run R1, we observe that most of the incorrect answers have only one supporting snippet. In two of the cases the problem comes from incorrect post-processing of the NER annotation produced by SIEMÊS. For example, in one case the correct answer was “primeiro imperador da China” and SIEMÊS only chunked “imperador da China”, which is correct from a NER point of view but is not enough for the answer. In such situations, RAPOSA would need to extract the additional information from SIEMÊS output, but it is not yet able to do it. In two other cases, the problem came from the inability of RAPOSA to choose the “best” answer from the set of candidates generated, all of them supported by the same number of snippets (only one). In some other cases, the supporting snippets obtained (again usually only one) are only slightly related to the topic of the

Table 1. Results of runs R1 and R2, with the corresponding number of questions where the context matching rules (cmr) or the simple type checking strategy (stcs) were applied

	R1 = 34 cmr				R2 = 34 cmr + 40 stcs				R2 - R1 = 40 stcs			
type	right	wrong	total	P_{type}	right	wrong	total	P_{type}	right	wrong	total	P_{type}
ANS	5	11	16	0.31	10	24	34	0.29	5	13	18	0.28
NUL1	1	14	15	0.07	7	29	36	0.19	6	15	21	0.29
NUL2	0	3	3	0	0	4	4	0	0	1	1	0
global	6	28	34	0.18	17	57	74	0.23	11	29	40	0.28

question, and are actually misleading for the context analysis rules, which are still very naive. One of the most important causes for this problem is related to the rudimentary stemming procedure implemented in the Query Generator. Better query expansion techniques are obviously needed.

The global precision achieved in R2, in which more than 50% of the questions attempted were addressed using the *simple type checking strategy*, was 0.23. If we analyze the answers from run R2 in more detail it is possible to see that about 20% of the incorrect answers result from incorrectly choosing the answer candidate among those elements in the snippet that in fact have the correct admissible semantic type. For example, RAPOSA answered “Lisboa” to the question “Onde morreu Kurt Cobain?” / “Where did Kurt Cobain die?” because there was a reference to Seattle and to Lisboa in the snippet extracted. Another very frequent type of error in run R2, which was not so evident in R1, reveals an important weakness of the Snippet Searcher. After querying the database based on keywords, and after tagging the text using SIEMÊS, the Snippet Searcher does not check whether keywords in the sentence match the expected semantic type of the keywords in the query. This may lead to many problematic situations if keywords have frequent homographs. This is especially severe for questions involving places or people because it is very common for people to have surnames that are also place names. A recent study [13] reports that over 30% of the names of people found in a 32k document sample of the Portuguese web contained at least one geographic / place name. This figure alone illustrates the magnitude of this particular problem.

5 Further Work

We still have to go a long way until RAPOSA reaches the performance level required for generating biographies using an iterative question-answering strategy. Future work will necessarily concentrate on improving existing modules of RAPOSA, especially on solving the various problems already identified. The overall performance of RAPOSA will greatly benefit from more efficient query expansion techniques and from a proper answer fusion module, taking into account biography-related issues. Some work will also be done on improving better context matching rules. The focus of our work will continue to be developing efficient means for answering specific biography-related questions.

Acknowledgements. This work was partially supported by grants SFRH / BD / 23590 / 2005 and POSI / PLP / 43931 / 2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

References

1. Feng, D., Ravichandran, D., Hovy, E.H.: Mining and Re-ranking for Answering Biographical Queries on the Web. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006), Boston, Massachusetts, USA, pp. 1283–1288. AAAI Press, Stanford, California (2006)

2. Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C.: Priberam's question answering system for Portuguese. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 410–419. Springer, Heidelberg (2006)
3. Quaresma, P., Rodrigues, I.: A logic programming based approach to the QA@CLEF'05 track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 351–360. Springer, Heidelberg (2006)
4. Sarmento, L.: SIEMES - a named entity recognizer for Portuguese relying on similarity rules. In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 90–99. Springer, Heidelberg (2006)
5. Mota, C., Santos, D., Ranchhod, E.: Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. In: AVALON, pp. 161–175.
6. Srihari, R.K., Li, W.: A Question Answering System Supported by Information Extraction. In: Proceedings of the 6th Conference on Applied Natural Language Processing, pp. 166–172. Morgan Kaufmann Publishers, San Francisco (2000)
7. Moldovan, D.I., Harabagiu, S.M., Girju, R., Morarescu, P., Lacatusu, V.F., Novischi, A., Badulescu, A., Bolohan, O.: LCC Tools for Question Answering. In: NIST (ed.) Proceedings of the 11th Text REtrieval Conference (TREC-2002), November 19–22, Gaithersburg, MD, pp. 144–155 (2002)
8. Aunimo, L., Kuuskoski, R.: Question Answering using Semantic Annotation. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 477–487. Springer, Heidelberg (2006)
9. Costa, L.: Esfinge - a modular question answering system for Portuguese. In: CLEF2006WN (2006)
10. Sarmento, L.: Hunting answers with RAPOSA (FOX). In: CLEF2006WN (2006)
11. Sarmento, L.: BACO - A large database of text and co-occurrences. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., Tapias, D. (eds.) Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), Genoa, Italy, May 22–28, 2006, pp. 1787–1790 (2006)
12. Cardoso, N., Martins, B., Gomes, D., Silva, M.J.: WPT 03: a primeira coleção pública proveniente de uma recolha da web portuguesa. In: AVALON, pp. 279–288 (2003)
13. Chaves, M.S., Santos, D.: What Kinds of Geographical Information Are There in the Portuguese Web? In: Vieira, R., Quaresma, P., Nunes, M.d.G.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.) PROPOR 2006. LNCS (LNAI), vol. 3960, pp. 264–267. Springer, Heidelberg (2006)
14. Peters, C., Gey, F., Gonzalo, J., Müller, H., Jones, G.J., Kluck, M., Magnini, B., de Rijke, M. (eds.): Accessing Multilingual information Repositories. In: CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
15. Vieira, R., Quaresma, P., Nunes, M.V., Mamede, N.J., Oliveira, C., Dias, M.C. (eds.): PROPOR 2006. LNCS (LNAI), vol. 3960. Springer, Heidelberg (2006)
16. Nardi, A., Peters, C., Vicedo, J.L.: Working Notes of the Cross-Language Evaluation Forum Workshop, 20–22 September, Alicante, Spain (2006)
17. Santos, D.: Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. IST Press (2007)

The Effect of Entity Recognition on Answer Validation

Álvaro Rodrigo, Anselmo Peñas, Jesús Herrera, and Felisa Verdejo

Departamento de Lenguajes y Sistemas Informáticos

Universidad Nacional de Educación a Distancia

Madrid, Spain

{alvaroroy, anselmo, jesus.herrera, felisa}@lsi.uned.es

Abstract. The Answer Validation Exercise (AVE) 2006 is aimed at evaluating systems able to decide whether the responses of a Question Answering (QA) system are correct or not. Since most of the questions and answers contain entities, the use of a textual entailment relation between entities is studied here for the task of Answer Validation. We present some experiments concluding that the entity entailment relation is a feature that improves a SVM based classifier close to the best result in AVE 2006.

1 Introduction

The Answer Validation Exercise (AVE) [8] of the Cross Language Evaluation Forum (CLEF) 2006 is aimed at developing systems able to decide whether the responses of a Question Answering (QA) system are correct or not using a Textual Entailment approach [3]. The AVE 2006 test corpus contains hypothesis-text pairs where the hypotheses are built from the questions and answers of the real QA systems, and the texts are the snippets given by the systems to support their answers. Participant systems in AVE 2006 must return a value YES or NO for each hypothesis-text pair to indicate if the text entails the hypothesis or not (i.e. the answer is correct according to the text).

The questions and answers of the QA Track at CLEF contain many entities (person names, locations, numbers, dates...) due to the nature of the questions in this evaluation: 75% of the questions in Spanish were factoids in the previous campaign [10]. For this reason, we studied the consideration of an entity entailment relation in combination with a SVM classifier to solve the Answer Validation problem.

Section 2 describes the process of entity detection and entailment decision. Section 3 describes the experimental setting combining the entity based entailment decision with a baseline SVM system. Section 4 shows the results of the experiments together with an error analysis. Finally, we give some conclusions and future work in Section 5.

2 Entailment Between Entities

The first step for considering an entailment relation between entities is to detect them in a robust way that minimizes the effect of errors in the annotation.

The second step is the definition and implementation of an entailment relation between entities (numeric expressions, temporal expressions and named entities in our case).

Next subchapters describe in detail these steps involved in the decision of entity entailment.

2.1 Entity Recognition

We have used the FreeLing¹ [2] Name Entity Recognizer (NER) to tag numeric expressions (NUMEX), named entities (NE) and time expressions (TIMEX) of both text and hypothesis. The first approach was to use the categories given by the tool. However, the ambiguity of these expressions was not always solved correctly by the tool. Figure 1 shows an example of this problem. The expression 1990 is a year but it is recognized as a numeric expression in the hypothesis. However, the same expression is recognized as a temporal expression in the text and, therefore, the expression in the hypothesis cannot be entailed by it. This kind of errors that are not consistent between texts and hypotheses, breaks the possible entailment relation between entities.

This fact led us to ignore the entity categorization given by the tool and assume that text and hypothesis are related and close texts where same expressions must receive same categories, without the need of disambiguation. Thus, all detected entities received the same tag.

```
<t>...Irak invadió Kuwait en <TIMEX>agosto_de_1990</TIMEX>...</t>
<h>Irak invadió Kuwait en <NUMEX>1990</NUMEX></h>
```

Fig. 1. Example of error disambiguating the entity type

2.2 Entailment Relations Between Entities

Once the entities of the hypothesis and the text are detected, the next step is to determine the entailment relations between the entities in the text and the entities in the hypothesis.

We defined the following entailment relations between entities.

1. A Named Entity E_1 entails a Named Entity E_2 if the text string of E_1 contains the text string of E_2 . For example, *Yaser_Arafat* entails *Arafat*, but *Arafat* does not entail *Yaser_Arafat* (see example in Figure 2). Sometimes some characters change in different expressions of the same entity as, for example, in a proper noun with different wordings (e.g. *Yasser*, *Yaser*, *Yasir*). To detect the entailment in these situations, when the previous process fails, we implemented the edit distance of Levenshtein [7] following [6]. Thus, if two Named Entities differ in less than 20%, then we assume that there exists an entailment relation between these entities.
2. A Time Expression T_1 entails a Time Expression T_2 if the time range of T_1 is included in the time range of T_2 . For example, *26 April 1986* entails *April 1986* but not in the other sense. The first approach to implement this is to consider time

¹ <http://garraf.epsevg.upc.es/freeling/>

expressions as strings and check if one string contains the other in a similar way to the Named Entities. However, there are several ways to express the same date with different text strings. The second approach is to consider the normalized form of the time expressions. However, the normalization depends on the correct detection of the entity type. Thus the best approach is to give the two chances for the entailment, the original string and a normalized form. Figure 2 shows a positive example of entailment between temporal expressions.

3. A numeric expression N_1 entails a numeric expression N_2 if the range associated to N_2 encloses the range of N_1 . In [5] we considered the units of the numerical expressions where the unit of N_1 must entail the unit of N_2 . For example, *17,000,000 citizen* entails *more than 15 million people*. However, sometimes the units detection needs some kind of anaphora resolution and we ignored the units for the experiments here described. Figure 2 shows an example of numeric expression where the normalization is considered allowing the detection of the entailment relation.

```
<t>...presidida por <ENTITY>Yaser_Arafat</ENTITY>...</t>
<h><ENTITY>Arafat</ENTITY> preside la ...</h>

<t>...durante la fuga de <ENTITY>Chernobyl</ENTITY> el
    <ENTITY>26_de_abril_de_1986</ENTITY>...</t>
<h>La catástrofe de <ENTITY>Chernobyl</ENTITY> ocurrió en
    <ENTITY>abril_de_1986</ENTITY></h>

<t>Los seis países citados forman parte
    de los <ENTITY>diez</ENTITY> europeos que ingresaron ya... </t>
<h>...está formada por <ENTITY>10</ENTITY> países.</h>
```

Fig. 2. Pairs that justify the process of entailment

3 Experimental Setting

The system developed for the AVE 2006 in Spanish is based on the ones developed for the First [3] and the Second [1] Recognizing Textual Entailment (RTE) Challenges for English.

In the system here described, the basic ideas from the ones presented to the RTE Challenges were kept, but the new system was designed and developed according to the available resources for the Spanish language, lacking some subsystems implemented in English such, for example, dependency analysis.

The proposed system is based on lemmatization. The system accepts pairs of text snippets (text and hypothesis) at the input and gives a boolean value at the output: YES if the text entails the hypothesis and NO otherwise. This value is obtained by the application of a learned model by a SVM classifier. System components, are the following:

- *Linguistic processing*: The lemmas of every text and hypothesis pairs are obtained using Freeling.

- *Sentence level matching*: A plain text matching module calculates the percentage of words, unigrams (lemmas), bigrams (lemmas) and trigrams (lemmas), respectively, from the hypothesis entailed by lexical units (words or n-grams) from the text, considering them as bags of lexical units.
- *Entailment Decision*: A SVM classifier, from Yet Another Learning Environment (Yale 3.0) [4], was applied in order to train a model from the development corpus given by the organization and to apply it to the test corpus.

3.1 SVM with Baseline Attributes

The SVM model was trained by means of a set of features obtained from the sentence level matching module for every pair <text, hypothesis>:

1. Percentage of words of the hypothesis present in the text (treated as bags of words).
2. Percentage of unigrams (lemmas) of the hypothesis present in the text (treated as bags of unigrams).
3. Percentage of bigrams (lemmas) of the hypothesis present in the text (treated as bags of bigrams).
4. Percentage of trigrams (lemmas) of the hypothesis present in the text (treated as bags of trigrams).

The first experiment evaluated the results obtained by this baseline system.

3.2 Entailment Decision Based Only on Entities Entailment

The thesis we follow in the recognition of textual entailment is that all the elements in the hypothesis must be entailed by elements of the supporting text. In special, all the entities in the hypothesis must be entailed by entities in the supporting text. Therefore, the system assumes that if there is an entity in the hypothesis not entailed by one or more entities in the text, then the answer is not supported and the system must return the value NO for that pair.

However, in pairs where all the entities in the hypothesis are entailed, there is not enough evidence to decide if the value of the pair is YES or NO.

In this experiment we decided to evaluate the results obtained when the default value is always YES except if there exist entities not entailed in the hypothesis (evidence of a not supported answer).

3.3 Check Entities Entailment Before SVM Classification

As we argued above (section 3.2), the information about entities is useful to detect some pairs without entailment. However, there is not enough evidence to decide the entailment value in the pairs where all the entities in the hypothesis are entailed. Therefore, a solution is to apply the SVM setting (section 3.1) only in this case. First, the system assigns the value NO to the pairs with entities not entailed in the hypothesis and second, the rest of pairs are classified with the baseline SVM system.

3.4 SVM Classification Adding the Attribute of Entity Entailment

The last experiment was to use the information of entailment between entities as an additional attribute in the SVM classifier, in a similar way we already tested in English with numeric expressions in the system of the RTE2 Challenge [5].

4 Results and Error Analysis

The proposed system has been tested over the Spanish test set of the AVE 2006. Table 1 shows the precision, recall and F-measure for the different settings compared with a baseline system that always returns YES and the system that obtained the best results at AVE 2006 (COGEX) [9].

Table 1. Results of the experiments compared with the best and the baseline systems

System	F-measure	Precision over YES	Recall over YES
Best AVE 2006 System	0.61	0.53	0.71
SVM classification adding the attribute of entity entailment	0.60	0.49	0.77
Entailment decision based only on entities entailment	0.59	0.46	0.83
SVM with baseline attributes	0.56	0.47	0.71
Check entities entailment before SVM classification	0.55	0.46	0.71
100% YES Baseline	0.45	0.29	1

The system based on SVM with the basic attributes obtained an F of 0.56 which is not bad compared with the baseline (F of 0.45). However, this result is worse than considering only the entailment relation between entities (F of 0.59). This is due to the higher recall obtained (0.83) after giving a YES value to the pairs that passed the entity entailment test. As it was mentioned before, the answer validation decision based only on the entailment between entities has a good performance for detecting the wrong answers (achieving an 89% of precision in the detection of pairs without entailment), but it does not provide enough evidence for most of the pairs. For this reason we used the system based on SVM with the basic attributes to decide on the rest of pairs. However, this setting obtained worse results (F of 0.55). The appropriate configuration was to include the entailment relation between entities as an additional attribute inside the SVM, obtaining an F of 0.60, close to the 0.61 obtained by COGEX.

Most of the errors in the entailment relation between entities were due to a wrong detection of the entities. In many cases the QA systems changed the original text, ignoring the writing of the named entities and returning all the words of the supporting snippets in lower case. In these cases, the NER tool cannot recognize any named entity in the text and the entities in the hypothesis are not entailed.

In some other cases the answers are given in capital letters and the NER tool recognize all words as named entities despite of they are not usually named entities. Therefore there are some false entities in the hypothesis that cannot be entailed by the text.

To solve these problems we need more robust NER tools and, at same time, the QA systems must take into account that their output can be taken as input by answer validation systems. QA systems should return their responses as they appear in the original texts.

5 Conclusions and Future Work

We have defined an entailment relation between entities and we have tested its use in an Answer Validation system. The addition of this relation as an additional attribute in a SVM setting improves the results of the system close to the best results in the AVE 2006.

Compared with the best in AVE 2006, our system gets higher recall and lower precision suggesting that we still have room to work on more restrictive filters to detect pairs without entailment.

Future work is oriented to the development of more robust NER tools and richer implementations of the entailment relations between entities.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Science and Technology within the project R2D2-SyEMBRA. (TIC-2003-07158-C04-02), the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and a PhD grant by UNED.

References

1. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The Second PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venezia, Italy (April 2006)
2. Carreras, X., Chao, I., Padró, L., Padró, M.: FreeLing: An Open-Source Suite of Language Analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal (2004)
3. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, UK, April 2005, pp. 1–8 (2005)
4. Fischer, S., Klinkenberg, R., Mierswa, I., Ritthoff, O.: Yale 3.0, Yet Another Learning Environment. User Guide, Operator Reference, Developer Tutorial. Technical report, University of Dortmund, Department of Computer Science. Dortmund, Germany (2005)
5. Herrera, J., Peñas, A., Rodrigo, Á., Verdejo, F.: UNED at PASCAL RTE-2 Challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venezia, Italy, April 2006, pp. 38–43 (2006)
6. Herrera, J., Peñas, A., Verdejo, F.: Textual entailment recognition based on dependency analysis and WordNet. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 231–239. Springer, Heidelberg (2006)

7. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics - Doklady* 10, 707–710 (1966)
8. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F. (eds.): Overview of the answer validation exercise 2006. LNCS. Springer, Heidelberg (2007)
9. Tatu, M., Iles, B., Moldovan, D.: Automatic Answer Validation using COGEX. In: Working Notes. CLEF 2006 Workshop, September 20-22, Alicante, Spain (2006)
10. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 Multilingual Question Answering Track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 307–331. Springer, Heidelberg (2006)

A Knowledge-Based Textual Entailment Approach Applied to the AVE Task*

Ó. Ferrández, R.M. Terol, R. Muñoz, P. Martínez-Barco, and M. Palomar

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{ofe,rafamt,rafael,patricio,mpalomar}@dlsi.ua.es

Abstract. We propose a system that has been initially used for Recognising Textual Entailment (RTE). Due to the fact that these two tasks (AVE and RTE) rely on the same main idea, our system was considered appropriate to be applied to AVE. Our system represents texts by means of logic forms and computes the semantic similarity between them. We have also designed a voting strategy between our system and the MLEnt system also developed at the University of Alicante. Although the results have not been very high, we consider them quite promising.

1 Introduction

In our participation in AVE, we want to evaluate the positive impact that our system can produce in the context of QA. Initially, our system [5] was developed and evaluated for Recognising Textual Entailment (RTE) [1] in English language. AVE and RTE basically have the same purpose, that is to find semantic implications between two fragments of text, and for this reason our system was appropriate for the AVE competition.

The rest of this paper is organized as follows. The next section presents the description of our system and its components. Section 3 illustrates the experiments and the results. Finally, section 4 wraps up the paper with conclusions and future work.

2 System Description

As we have mentioned in the previous section, the system that we describe here has already been used to solve Textual Entailment. A detailed description of our system is depicted in [5]. In this paper, we only make a brief overview of the entire system and of its two main components.

The system works according to the followings steps:

* This research has been partially funded by the Spanish Government under project CICYT number TIC2003-07158-C04-01.

1. It obtains the logic forms of the two given snippets.
2. It computes the semantic similarity between the logic forms. This step will provide a semantic weight that will determine a true or false entailment.
3. It compares the semantic weight obtained in the previous step to an empiric threshold acquired from the development corpus. If this weight is greater than the threshold, then the entailment holds.

2.1 Derivation of the Logic Forms

A logic form can be defined as a set of predicates related among them which have been inferred from a sentence. The aim of using logic forms is to simplify the sentence treatment process. In our approach, we use a format for representing logic forms similar to the format of the lexical resource called Logic Form Transformation of eXtended WordNet (LFT) [2]. And the process to infer the logic form associated of a sentence is through applying NLP rules to the dependency relationship of the words. We use MINIPAR¹ in order to obtain these dependency relationships. Once the dependency relationships have been acquired, the next step is the analysis of these dependencies by means of several NLP rules that transform the dependency tree into its logic form associated.

2.2 Computation of Similarity Measures

The main idea is that the verbs generally govern the meaning of sentences. For this reason, this component is initially focused on analysing semantic relations between the verbs of the two logic forms derived. And secondly, if there is a relation between the verbs, then the method will analyse the similarity relations between all predicates depending on the two verbs. In the case of there is not semantic relations, the method will not analyse any more logic form predicate.

In order to obtain the semantic similarity between the predicates of the logic forms, two approaches have been implemented:

- **Based on WordNet relations:** we determine if two predicates are related through the composition of the WordNet relations. We consider relations of hyponymy, entailment and synonymy. And, if there is a path which connects these two predicates by means of these relations, we conclude that these predicates are semantically related with a specific weight. The length of the path must be lower or equal than 4. Each WordNet relation has a weight attached to it, and the final weight is calculated as the product of the weights associated to the relations connecting the two predicates. The assigned weights and the entire module is described in detail in [6].
- **Based on Lin's measure [4]:** in this case, the similarities were computed using the Lin's measure as is implemented in WordNet::Similarity².

Finally, we want to point out that a Word Sense Disambiguation module has not been employed in deriving the WordNet relations between any two predicates. Only the first 50% of the WordNet senses were taken into account.

¹ <http://www.cs.ualberta.ca/~lindek/minipar.htm>

² <http://www.d.umn.edu/~tpederse/similarity.html>

2.3 UA-Voting

As the University of Alicante has two systems based on different techniques which solve Textual Entailment. We want to evaluate each system individually in the innovative AVE task as well as to check whether a combination of these two systems improves the results. The two systems involved are: the system described in this paper and the system presented in [3], called MLEnt.

A run was obtained by combining the output of the two systems. The system were run only for English and their output was merged using a voting strategy. The final output was derived from three different runs of the two systems: our system's output employing the semantic similarity module based on Lin's measure and two MLEnt outputs resulted from two different experiments with skip-grams and the longest common subsequence technique³.

3 Results and Discussion

For the development and the evaluation of our system, we used the corpus provided by the AVE organizers. The development corpus for English has around 2870 pairs test-hypothesis, but only 168 are revised manually. We only used the revised pairs in order to adjust our system for the AVE task. The test data contains 2088 pairs. Table 1 illustrates the results achieved by our two semantic similarity approaches individually and regarding UA-voting experiment.

Table 1. AVE 2006 official results for English language

Development data	Precision YES pairs	Recall YES pairs	F-measure
WNrelations	0.2368	0.75	0.36
Lin	0.2265	0.8055	0.3536
Test data	Precision YES pairs	Recall YES pairs	F-measure
WNrelations	0.2072	0.5116	0.2949
Lin	0.1981	0.6884	0.3077
UA-voting	0.2054	0.6047	0.3066

As we can observe in Table 1, all the results are quite similar with respect to F-measure. Using the approach based on Lin's measure our system achieved better recall. However, these differences are insignificant when it comes to deciding what approach works better. The run corresponding to the combination of the two systems developed at the University of Alicante did not achieve the expected results. This fact proves that we have to investigate other voting strategies or, perhaps other ways of joining the two different systems' technologies in order to obtain only one system.

³ For further details see [3].

4 Conclusions and Future Work

In this paper, we have presented a system based on logic forms and semantic comparison between them using two different approaches based on WordNet relations and Lin's semantic measure. Moreover, we also present a voting strategy combining the two systems developed at the University of Alicante.

The results obtained have not been very high, but quite promising. We want to attach great importance to the fact that, in the RTE-2 Challenge [1] we achieved 60% in avg. precision, but in the case of AVE the results have decreased dramatically. This supports the claim that research in the area of textual entailment is still in its very early stages therefore, there is a long way to go.

As future work, we want to investigate in depth the corpus provided by AVE, as our system was initially developed for dealing with entailment relations belonging to many NLP applications (see the RTE2 specification [1]) such as Information Retrieval, Text Summarization, etc. This detailed investigation could yield improvements specific to the AVE task. Possibly, in order to solve some deficiencies, we need to improve our method by investigating in more detail the dependency trees of the text and the hypothesis and how the addition of other NLP tools such as a Named Entity Recognizer could help in detecting entailment relations. Finally, with this kind of knowledge we will be able to integrate our system within a module performing answer validation for QA.

References

1. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The Second PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, RTE2 2006, pp. 1–9 (2006)
2. Harabagiu, S., Miller, G.A., Moldovan, D.I.: WordNet 2 - A Morphologically and Semantically Enhanced Resource. In: Proceedings of ACL-SIGLEX99: Standardizing Lexical Resources, Maryland, pp. 1–8 (1999)
3. Kozareva, Z., Montoyo, A.: MLEnt: The Machine Learning Entailment System of the University of Alicante. In: Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, RTE2 2006, pp. 16–21 (2006)
4. Lin, D.: An Information-Theoretic Definition of Similarity. In: ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304. Morgan Kaufmann, San Francisco (1998)
5. Ferrández, Ó., Terol, R.M., Muñoz, R., Martínez-Barco, P., Palomar, M.: An Aporach based on Logic Forms and WordNet relationships to Textual Entailment Performance. In: Proceedings of the Second PASCAL Recognising Textual Entailment Challenge, RTE2 2006, pp. 22–26 (2006)
6. Ferrández, Ó., Terol, R.M., Muñoz, R., Martínez-Barco, P., Palomar, M.: Deep vs. Shallow Semantic Analysis Applied to Textual Entailment Recognition. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.) FinTAL 2006. LNCS (LNAI), vol. 4139, pp. 225–236. Springer, Heidelberg (2006)

Automatic Answer Validation Using COGEX

Marta Tatu, Brandon Iles, and Dan Moldovan

Language Computer Corporation, Richardson, Texas, 75080, USA
{marta,biles,moldovan}@languagecomputer.com
www.languagecomputer.com

Abstract. This paper reports the performance of Language Computer Corporation's natural language logic prover for the English and Spanish subtasks of the Answer Validation Exercise. COGEX takes as input a pair of plain English text snippets, transforms them into semantically rich logic forms, automatically generates natural language axioms which will be used during the search for a proof, and determines the degree of entailment between the entailing text snippet and a possible relaxed version of the entailed text. The system labels an input pair as positive entailment if its proof score is above the threshold learned during training. Our semantic logic-based approach achieves 43.93% F-measure for the English data and 60.63% on Spanish.

1 Introduction

While communicating, humans use different expressions to convey the same meaning. One of the challenges for natural language understanding systems is to determine if the meaning of one text can be derived from the meaning of another. A module that recognizes the *textual entailment* between two snippets can be employed within Question Answering (QA) settings. A subsystem designed to identify semantic entailments can validate the correctness of the answers returned. This idea was promoted as the Answer Validation Exercise (AVE)¹ subtask of the 2006 QA track at the Cross-Language Evaluation Forum.

In this paper, we describe a model to represent the knowledge encoded in text as well as a logical setting suitable for a semantic entailment recognition system. We cast the textual inference problem as a *logical implication between meanings*. The entailing supportive text (T) semantically infers the entailed hypothesis (H), if its meaning logically implies the meaning of H . Thus, our system first transforms both text fragments into logic forms and captures their meaning by detecting the *semantic* relations between the constituents of T and between the constituents of H . It then loads these rich logic representations into a natural language logic prover to decide whether the entailment holds and subsequently provides a justification for it. This approach [13] proved to be highly effective at the Second PASCAL Recognizing Textual Entailment Challenge [2]. Figure 1 illustrates our answer validation system's architecture. We used this setup for

¹ <http://nlp.uned.es/QA/ave/index.php>

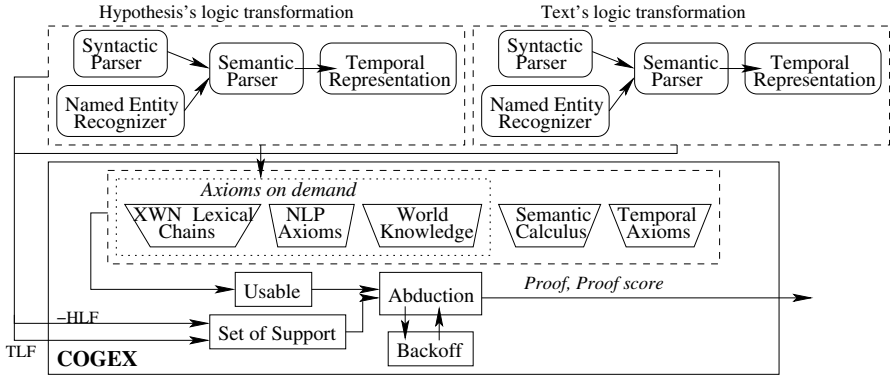


Fig. 1. COGEX's Architecture

the English dataset and for the automatic translation (in English) of the Spanish data². LCC's machine translation system implements a phrase-based statistical translation method [6]. It was trained using the Europarl corpus [5]. The details of our translation solution is described in [12]. To our knowledge, there are few logical approaches to textual entailment. [3] represents T and H into a first-order logic translation of the DRS language used in Discourse Representation Theory and uses a theorem prover and a model builder with some generic, lexical, and geographic background knowledge to prove the entailment between the two texts. [4] proposes a Description Logic-based knowledge representation language used to induce the representations of T and H and uses an extended subsumption algorithm to check if any of T 's representations, obtained through equivalent transformations, entails H .

1.1 Cogex – A Logic Prover for NLP

LCC's entailment system uses COGEX [9], a natural language prover originating from OTTER [7]. For this *recognizing textual entailment* task, the negated form of the hypothesis ($\neg H$) and T 's predicates form the *set of support* list. For the *usable* list, we consider several types of axioms: eXtended WordNet, linguistic, semantic, and temporal axioms. Once the set of support and usable lists are created, the logic prover begins its search for a proof. When it finds a contradiction, the proof is complete. If a refutation cannot be found, then arguments are relaxed. If this fails, entire predicates are dropped from H . The goal of the iterative relaxation process is to allow the prover to output (partial) proofs when the provided background knowledge is incomplete and to compensate for parsing errors.

Once a proof by contradiction is found, its score is computed by starting with an initial perfect score and deducting points for each axiom utilized in the proof, each relaxed argument, and each dropped predicate. The computed

² Any processing of the Spanish data was performed on its English translation.

score is a measure of the kinds of axioms used in the proof and the significance of the dropped arguments and predicates. Due to the logic prover's relaxation techniques, it is always successful in producing a proof. The determination of whether entailment exists is made by examining the normalized value of the proof's score³. If this final proof score is above a threshold learned on the development data, then the pair is labeled as positive entailment.

2 Knowledge Representation

For the textual entailment task, our logic prover uses a dual-layered logical representation which captures the text's syntactic and semantic propositions.

2.1 Logic Form Transformation

In the first stage of our entailment process, COGEX converts T and H into logic forms [11]. This first syntactic layer of the logic representation is automatically derived from a full parse tree and acknowledges syntax-based relationships. If we consider, for instance, the hypothesis *Iraq invaded the country of Kuwait in 1990*, `iraq_NN(x1) & _country_NE(x1) & invade_VB(e1,x1,x2) & country_NN(x2) & of_IN(x2,x3) & kuwait_NN(x3) & _country_NE(x3) & in_IN(x3,x4) & 1990_NN(x4) & _date_NE(x4)` constitutes its first layer of logic representation. Exceptions to the one-predicate-per-open-class-word rule include the adverbs *not* and *never*. In cases similar to *the provisions of the convention do not affect in any way the exercise of navigational rights and freedoms*, the system removes `not_RB(x3,e1)` and negates the verb's predicate (`-affect_VB(e1,x1,x2)`) unless the verb is modified by an adverb (*They don't admit it in public* is represented as `admit_VB(e1,x8,x9) & -in_public_RB(x11,e1)`).

2.2 Semantic Relations

The second layer of our logic representation adds the semantic relations, the underlying relationships between concepts. Our representation module maps each semantic relation, identified by LCC's semantic parser, to a predicate whose arguments are the events and entities that participate in the relation. It then adds these semantic predicates to the logic form. For example, the logic form of *Iraq invaded the country of Kuwait in 1990* is augmented with the `AGENT_SR(x1,e1) & THEME_SR(x2,e1) & ISA_SR(x3,x2) & TIME_SR(x4,e1)` relations⁴ (*Iraq* is the *agent* of the *invade* event, *the country of Kuwait* is its *theme*, etc.).

2.3 Temporal Representation

In addition to semantic predicates, we represent every *date/time* into a normalized form `time_TMP( BeginFn(xi), startYear, startMonth, startDay,`

³ Because (T, H) pairs with longer sentences can potentially drop more predicates and receive a lower score, COGEX normalizes the proof scores by dividing the assessed penalty by the maximum assessable penalty (all the predicates from H are dropped).

⁴ $R(x, y)$ should be read as " x is R of y ".

startHour, startMinute, startSecond) & time_TMP(EndFn(xi), endYear, endMonth, endDay, endHour, endMinute, endSecond). Furthermore, temporal reasoning predicates are derived from both the detected semantic relations and a module which utilizes a learning algorithm to detect temporally ordered events ($(S \text{ sequence}, E_1, E_2) \equiv \text{earlier\_TMP}(e1, e2)$, $(S \text{ contain}, E_1, E_2) \equiv \text{during\_TMP}(e1, e2)$, etc.).

3 Axioms on Demand

COGEX generates axioms on demand for a given (T, H) pair whenever the semantic connectivity between two concepts needs to be established in a proof.

3.1 eXtended WordNet Lexical Chains

For the semantic entailment task, the ability to recognize two semantically-related words is an important requirement. Therefore, we automatically construct lexical chains of WordNet relations from T 's constituents to H 's constituents [10]. We use the first k senses for each word⁵, unless the source and the target of the chain are synonyms. If a chain exists, the system generates an axiom from T 's predicates to H 's. For example, given the ISA relation between *murder#1* and *kill#1*, the system generates the axiom $\text{murder\_VB}(e1, x1, x2) \rightarrow \text{kill\_VB}(e1, x1, x2)$.

LCC's lexical chain axioms append the entity name of the target concept, whenever it exists. For instance, when it tries to infer *electoral campaign is held in Nicaragua* from *Nicaraguan electoral campaign*, the logic prover uses the axiom $\text{Nicaraguan\_JJ}(x1, x2) \rightarrow \text{Nicaragua\_NN}(x1) \ \& \ \text{\_country\_NE}(x1)$.

We ensured the relevance of the lexical chains by limiting the maximum path length (to three relations) and the set of WordNet relations used to create the chains. For example, the automatic axiom generation module does not consider chains with an IS-A relation followed by a HYPONYMY link (*Chicago* $\xrightarrow{\text{is-a}}$ *city* $\xrightarrow{\text{hyponymy}}$ *Detroit*). Similarly, the system rejects chains with more than one HYPONYMY relations because the hypothesis should be more general than T and too many HYPONYMY relations can lead to a too specific concept in H . Another restriction imposed on the lexical chains generated for entailment is to not start from or include too general concepts⁶. Therefore, we assigned to each noun and verb synset from WordNet a generality weight. If d_i is the depth of concept c_i within its hierarchy H_i , D_{H_i} is the maximum depth and $IC(c_i) = -\log(p(c_i))$ is the information content of c_i measured on the British National Corpus, then

$$\text{generality}W(c_i) = \frac{1}{\frac{d_i+1}{D_{H_i}} * IC(c_i)} \quad (1)$$

⁵ Because WordNet senses are ranked based on their frequency, the correct sense is most likely among the first k . In our experiments, $k = 3$.

⁶ There are no restrictions on the target concept.

In our experiments, we discarded the chains with concepts whose generality weight exceeded 0.8, such as *object/NN/1*, *act/VB/1*, *be/VB/1*, etc.

Another important change that we introduced in our extension of WordNet is the refinement of the DERIVATION relation. Because of its ambiguity with respect to the role of the noun, we split this relation in three: ACT-DERIV, AGENT-DERIV and THEME-DERIV. For instance, the axioms $\text{act\_VB}(\mathbf{e1}, \mathbf{x1}, \mathbf{x2}) \rightarrow \text{acting\_NN}(\mathbf{e1})$ (ACT), $\text{act\_VB}(\mathbf{e1}, \mathbf{x1}, \mathbf{x2}) \rightarrow \text{actor\_NN}(\mathbf{x1})$ (AGENT) reflect different types of derivation.

3.2 NLP Axioms

Our NLP axioms are linguistic rewriting rules that help break down complex syntactic structures (complex nominals or coordinating conjunctions) and express syntactic equivalence (appositions). These axioms are made available only to the (T, H) pair that generated them. For example, the axiom $\text{nn\_NNC}(\mathbf{x3}, \mathbf{x1}, \mathbf{x2}) \ \& \ \text{Sigmund\_NN}(\mathbf{x1}) \ \& \ \text{Freud\_NN}(\mathbf{x2}) \rightarrow \text{Freud\_NN}(\mathbf{x3})$ breaks down the noun compound *Sigmund Freud* into *Sigmund* and *Freud*. This type of axiom proved to be the most frequently used for the English and Spanish datasets. The apposition axioms alone solved 33.22% of the Spanish entailments.

New subtypes of NLP axioms include implications from named entity classes to nouns that describe the class (for example, $\text{kuwait\_NN}(\mathbf{x1}) \ \& \ \text{\_country\_NE}(\mathbf{x1}) \rightarrow \text{country\_NN}(\mathbf{x1})$ helps COGEX infer *H's country of Kuwait* from texts which mention only *Kuwait*) and equivalences between name variations (for instance, *Mikhail Gorbachev*, *Michail Gorbatchov*, and *Mikhail S. Gorbachev*).

3.3 World Knowledge Axioms

Because the lexical or the syntactic knowledge can sometimes not solve an entailment pair, we exploit the WordNet glosses, an abundant source of world knowledge. We used the logic forms of the glosses provided by eXtended WordNet⁷ to automatically create our world knowledge axioms. We also incorporate in our system a small common-sense knowledge base of 457 hand-coded world knowledge axioms, 74 of which have been manually designed from the development set data, and 383 originating from previous projects. These axioms express knowledge that could not be derived from WordNet regarding employment (the axiom $\text{\_country\_NE}(\mathbf{x1}) \ \& \ \text{negotiator\_NN}(\mathbf{x2}) \ \& \ \text{nn\_NNC}(\mathbf{x3}, \mathbf{x1}, \mathbf{x2}) \rightarrow \text{work\_VB}(\mathbf{e1}, \mathbf{x2}, \mathbf{x4}) \ \& \ \text{for\_IN}(\mathbf{e1}, \mathbf{x1})$ helps the prover infer that *Christopher Hill works for the US* from *top US negotiator*, *Christopher Hill*), family relations, awards, etc.

4 Semantic Calculus

The Semantic Calculus axioms combine two semantic relations identified within a text fragment and increase the semantic connectivity of the text [14]. The

⁷ <http://xwn.hlt.utdallas.edu>

82 semantic axioms enable inference of unstated meaning from the semantics detected in text. For example, if T (*even further to include external factors such as weather conditions*) explicitly states $\text{THEME}(\text{external factors}, \text{include})$ and $\text{ISA}(\text{weather conditions}, \text{factors})$, the logic prover uses the $\text{ISA\_SR}(x1, x2) \ \& \ \text{THEME\_SR}(x2, e1) \rightarrow \text{THEME\_SR}(x1, e1)$ semantic axiom (the dominance of the THEME relation over ISA) to infer the $\text{THEME}(\text{weather conditions}, \text{include})$. Another frequent axiom is $\text{LOCATION\_SR}(x1, x2) \ \& \ \text{PARTWHOLE\_SR}(x2, x3) \rightarrow \text{LOCATION\_SR}(x1, x3)$. Given the text *John lives in Dallas, Texas* and using this axiom, the system infers that *John lives in Texas*. The system applies the 82 axioms independent of the concepts involved in the semantic composition.

5 Temporal Axioms

One of the types of temporal axioms that we load in our logic prover links specific dates to more general time intervals. For example, *August 1990* entails the year *1990*. This rule is used to prove that *Iraq invaded the country of Kuwait in 1990* from *Trade sanctions imposed on Iraq after it invaded Kuwait in August 1990*. These axioms are automatically generated before the proof search starts. Additionally, the prover uses a SUMO knowledge base of temporal reasoning axioms that consists of axioms for a representation of time points and time intervals, Allen [1] primitives, and temporal functions. For instance, *before* is a transitive primitive: $\text{before\_TMP}(e1, e2) \ \& \ \text{before\_TMP}(e2, e3) \rightarrow \text{before\_TMP}(e1, e3)$.

6 Experiments and Results

The AVE corpus consists of text-hypothesis pairs generated from the responses to the QA task. The text snippet returned by a QA system as supportive for its answer became the T and the question, transformed into a declarative sentence which includes the answer, became the hypothesis H . Table 1 present statistics about the datasets provided by the AVE organizers. The test pairs tagged as *unknown* (pairs with inexact answers or not judged by humans) were ignored in the performance evaluation of the system.

6.1 COGEX's Results

Table 2 summarizes COGEX's performance on the English and Spanish datasets. Because our knowledge representation method relies on a full syntactic parse

Table 1. Datasets Statistics

	Development (%)			Test (%)			
	True	False	Total	True	False	Unknown	Total
English	436 (15.19)	2434 (84.80)	2870	215 (10.29)	1144 (54.78)	729 (34.91)	2088
Spanish	638 (21.96)	2267 (78.03)	2905	671 (28.32)	1615 (68.17)	83 (3.50)	2369

Table 2. System’s performance on the English and Spanish datasets

Data (%)	Dev-Eng	Dev-Spa	Test-Eng	Test-Spa
Precision for YES class	34.49	53.58	31.87	52.70
Recall for YES class	72.93	78.52	70.70	71.39
F-measure for YES class	46.83	63.69	43.93	60.63
Accuracy	74.84	80.34	71.45	72.79

tree of the input and, in the AVE corpus, a large number of hypotheses were syntactically incorrect, the logic representations were compromised. We performed a deep analysis of the system’s results in order to pinpoint the sources of its mistakes. We felt there were pairs for which the system’s response correctly contradicted the human annotated label for the pair. For example, English pair 5183 with the text ... *still win the war. Gen. Dwight D. Eisenhower called D-day “a great crusade,” but privately, ...* and the hypothesis *EISENHOWER called D-day “a great crusade”*. One of the authors of this paper manually labeled the test pairs which were annotated by judges with either YES or NO (we discarded the UNKNOWN pairs) and we computed the inter-annotator agreement for this data. For the English YES/NO test pairs, we obtained a kappa agreement of 0.72 (0.92 proportional agreement) which shows substantial agreement and an almost perfect agreement for the Spanish data ($kappa = 0.88$ with 0.95 proportional agreement). These findings, combined with the system’s higher results for Spanish, suggest that the English data was more difficult or that pre-processing errors made the English pairs more problematic. Some of the system’s difficulties were posed by the semi-automatically generated English hypotheses, which contained the entire supporting text on the answer’s position. English test pair 6989 has, as *T*, *David Shearer wrote about Hodgkin, Bellany and Vermeer but failed to mention David Hosie ...* and, as *H*, *Vermeer was David Shearer wrote about Hodgkin, Bellany and Vermeer but failed to mention David Hosie* With so much overlap between *T* and *H*, the system assigned high scores to this type of pairs, which led to *true* assignments in 66.66% of false entailment cases (only 18.57% of the pairs with the entire *T* inserted in *H* are true entailments). On the other hand, the system effortlessly labelled as *false* the 34 negative pairs from the English test set with empty *T*s. For the Spanish data, 34 out of the 35 pairs whose *H* includes the whole *T* are false entailments and COGEX incorrectly labelled 41.17% of them.

Another subset of negative pairs that our system mistakenly tagged as positive includes pairs whose *T* contains the correct answer to the initial question, but another phrase mentioned in *T* was inserted in *H* on the answer position. For instance, *T*: *Jordan signed an accord with Israel on July 28 ending the state of war between them ...* is not entailing *H*: *Jordan and Israel signed a peace treaty on state of war* created from the question *On which day did Jordan and Israel sign a peace treaty?*, but it contains the correct answer. The high lexical overlap between *T* and *H* coupled with COGEX’s argument relaxation technique (the *on*-link between *signed* and *state of war* is relaxed) led to high proof scores.

7 Conclusion

In this paper, we presented a logic form representation of knowledge which captures syntactic dependencies, semantic relations between concepts, and temporal predicates. Several changes to our eXtended WordNet lexical chains module led to fewer unsound axioms. Semantic and temporal axioms decreased the prover's dependence on world knowledge. We plan to improve COGEX to detect false entailments even when the two texts have a high word overlap.

References

1. Allen, J.: Time and Time Again: The Many Ways to Represent Time. *International Journal of Intelligent Systems* 4(6), 341–355 (1991)
2. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The Second PASCAL Recognising Textual Entailment Challenge. In: *Proceedings of the Second PASCAL Challenges Workshop* (2006)
3. Bos, J., Markert, K.: Recognizing Textual Entailment with Logical Inference. In: *Proceedings of HLT/EMNLP 2005, Vancouver, Canada* (October 2005)
4. de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D., Sammons, M.: An Inference Model for Semantic Entailment in Natural Language. In: *Proceedings of AAAI-2005* (2005)
5. Koehn, P.: Europarl: A Parallel Corpus for Statistical MT. In: *MT Summit* (2005)
6. Koehn, P., Och, F.J., Marcu, D.: Statistical Phrase-based Translation. In: *Proceedings of HLT/NAACL, Edmonton, Canada* (2005)
7. McCune, W.W.: OTTER 3.0 Reference Manual and Guide (1994)
8. Moldovan, D., Clark, C., Harabagiu, S.: Temporal Context Representation and Reasoning. In: *Proceedings of IJCAI, Edinburgh, Scotland* (2005)
9. Moldovan, D., Clark, C., Harabagiu, S., Maiorano, S.: COGEX A Logic Prover for Question Answering. In: *Proceedings of the HLT/NAACL* (2003)
10. Moldovan, D., Novischi, A.: Lexical chains for Question Answering. In: *Proceedings of COLING, Taipei, Taiwan* (August 2002)
11. Moldovan, D., Rus, V.: Logic Form Transformation of WordNet and its Applicability to Question Answering. In: *Proceedings of ACL, France* (2001)
12. Olteanu, M., Suriyentrakorn, P., Moldovan, D.: Language Models and Reranking for Machine Translation. In: *Proceedings of the NAACL Workshop On Statistical Machine Translation* (2006)
13. Tatu, M., Iles, B., Slavick, J., Novischi, A., Moldovan, D.: COGEX at the Second Recognizing Textual Entailment Challenge. In: *Proceedings of the Second PASCAL Challenges Workshop* (2006)
14. Tatu, M., Moldovan, D.: A Semantic Approach to Recognizing Textual Entailment. In: *Proceedings of HLT/EMNLP* (2005)

Paraphrase Substitution for Recognizing Textual Entailment

Wauter Bosma¹ and Chris Callison-Burch²

¹ University of Twente,
NL-7500AE Enschede, The Netherlands
`bosmaw@cs.utwente.nl`

² University of Edinburgh,
2 Buccleuch Place, Edinburgh, EH8 9LW, United Kingdom
`callison-burch@ed.ac.uk`

Abstract. We describe a method for recognizing textual entailment that uses the length of the longest common subsequence (LCS) between two texts as its decision criterion. Rather than requiring strict word matching in the common subsequences, we perform a flexible match using automatically generated paraphrases. We find that the use of paraphrases over strict word matches represents an average F-measure improvement from 0.22 to 0.36 on the CLEF 2006 Answer Validation Exercise for 7 languages.

1 Introduction

Recognizing textual entailment has recently generated interest from a wide range of Natural Language Processing related research areas, such as automatic summarization, information extraction and question answering. Advances have been made with various techniques, such as aligning syntactic trees and word overlap. While there is still much room for improvement, Vanderwende and Dolan [1] showed that current approaches are close to hitting the boundaries of what is feasible with lexical-syntactic approaches.

Proposed directions to cross this boundary include using logical inference, background knowledge and paraphrasing [2]. We explore the possibility of applying paraphrasing to obtain a more reliable match between a given text and hypothesis for which the presence of an entailment relation is to be determined. For instance, consider the following RTE2 pair.

Text: *Clonaid said, Sunday, that the cloned baby, allegedly born to an American woman, and her family were going to return to the United States Monday, but where they live and further details were not released.*

Hypothesis: *Clonaid announced that mother and daughter would be returning to the US on Monday.*

In this example, text and hypothesis use different words to express the same meaning. Although deep inference is required to recognize that a mother is part

of the family, and that *daughter* and *baby* in this context most likely refer to the same person, most variation occurs on the surface level. For instance, *announced* in the hypothesis could be replaced by *said* without changing its meaning. Similarly, the phrases *the United States* and *the US* would not be matched by a system relying solely on word overlap.

The criterion that our system uses to decide whether a text entails a hypothesis is the length of the longest common subsequence (LCS) between the passages. Rather than identifying the LCS using word matching, our system employs an automatic paraphrasing method that extends matches to synonymous, but non-identical phrases. We automatically generate our paraphrases by extracting them from bilingual parallel corpora.

Whereas many systems use dependency parsers and other linguistic resources that are only available for a limited number of languages, our system employs a method that is comparatively language independent. For this paper we extract paraphrases in Dutch, English, French, German, Italian, Spanish, and Portuguese.

The paraphrase extraction algorithm is described in section 2. Section 3 describes how the entailment score is calculated, and how paraphrases are generated in the entailment detection system. The results of our participation in the CLEF2006 Answer Validation Exercise [3] (henceforth AVE) are presented in section 4. We will wrap up with a conclusion and directions for future work in section 5.

2 Paraphrase Extraction

Paraphrases are alternative ways of conveying the same information. The automatic generation of paraphrases has been the focus of a significant amount of research lately [4], [5], [6], [7]. In this work, we use Bannard and Callison-Burch's method [7], which extracts paraphrases from bilingual parallel corpora.

Bannard and Callison-Burch extract paraphrases from a parallel corpus by equating English phrases which share a common foreign language phrase. English phrases are aligned with their foreign translations using alignment techniques drawn from recent *phrase*-based approaches to statistical machine translation [8]. Paraphrases are identified by pivoting through phrases in another language.

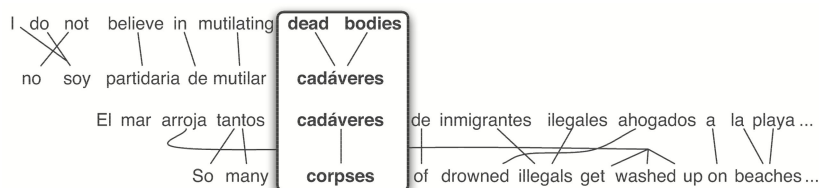


Fig. 1. A bilingual parallel corpus can be used to extract paraphrases

Table 1. Examples paraphrases and probabilities for the phrase *dead bodies*

bodies	0.21
dead bodies	0.17
body	0.09
deaths	0.07
dead	0.07
corpses	0.06
bodies of those killed	0.03
the dead	0.02
carcasses	0.02
corpse	0.01

Candidate paraphrases are found by first identifying all occurrences of the English phrase to be paraphrased, then finding its corresponding foreign language translations of the phrase, finally looking at what other English phrases those foreign languages translate back to.

Figure 1 illustrates how a Spanish phrase can be used as a point of identification for English paraphrases in this way. Often there are many possible paraphrases that can be extracted for a particular phrase. Table 1 shows the paraphrases that were automatically extracted for an English phrase. In order to assign a ranking to a set of possible paraphrases, Bannard and Callison-Burch define a paraphrase probability.

The paraphrase probability $p(e_2|e_1)$ is defined in terms of two translation model probabilities: $p(f|e_1)$, the probability that the original English phrase e_1 translates as a particular phrase f in the other language, and $p(e_2|f)$, the probability that the candidate paraphrase e_2 translates as the foreign language phrase. Since e_1 can translate as multiple foreign language phrases, f is marginalized out:

$$p(e_2|e_1) = \sum_f p(f|e_1)p(e_2|f) \quad (1)$$

The translation model probabilities can be computed using any standard formulation from phrase-based machine translation. For example, $p(e_2|f)$ can be calculated straightforwardly using maximum likelihood estimation by counting how often the phrases e and f were aligned in the parallel corpus:

$$p(e_2|f) \approx \frac{\text{count}(e_2, f)}{\sum_e \text{count}(e, f)} \quad (2)$$

We extend the definition of the paraphrase probability to include multiple corpora, as follows:

$$p(e_2|e_1) \approx \frac{\sum_{c \in C} \sum_f \text{in } c p(f|e_1)p(e_2|f)}{|C|} \quad (3)$$

Table 2. Examples of text/hypothesis pairs from CLEF AVE for which paraphrasing was required to make a correct assessment. The words in italics are hypothesis words which are aligned with the text sentence.

Pair: 573 (entailment: yes)
Text: Riding the crest of that wave is Latvia's new currency the handily named lat introduced two years ago.
Hypothesis: <i>The Latvian currency is lat.</i> (negative judgment; entailment score = 0.60)
Hypothesis paraphrase: <i>The Latvia currency lat.</i> (positive judgment; entailment score = 1.00)
Pair: 2430 (entailment: yes)
Text: India and Pakistan fought two of their three wars over control of Kashmir and their soldiers still face off across the Siachen Gracier 20 000 feet above sea level in the Himalayas.
Hypothesis: <i>India and Pakistan have fought two wars</i> for the possession of <i>Kashmir</i> . (negative judgment; entailment score = 0.67)
Hypothesis paraphrase: <i>India and Pakistan fought two of their wars</i> for possession of <i>Kashmir</i> . (positive judgment; entailment score = 0.83)
Pair: 8597 (entailment: yes)
Text: Anthony Busuttil, Professor of Forensic Medicine at Edinburgh University, examined the boy.
Hypothesis: <i>Anthony Busuttil is professor of Forensic Medicine at the University of Edinburgh.</i> (negative judgment; entailment score = 0.67)
Hypothesis paraphrase: <i>Anthony Busuttil professor of Forensic Medicine at Edinburgh University.</i> (positive judgment; entailment score = 1.00)

where c is a parallel corpus from a set of parallel corpora C . Thus multiple corpora may be used by summing over all paraphrase probabilities calculated from a single corpus (as in Equation 1) and normalizing by the number of parallel corpora. We calculate the paraphrase probabilities using the Europarl parallel corpus [9], which contains parallel corpora for Danish, Dutch, English, French, Finnish, German, Greek, Italian, Portuguese, Spanish and Swedish.

The method is multilingual, since it can be applied to any language which has a parallel corpus. Thus paraphrases can be easily generated for each of the languages in the CLEF AVE task using the Europarl corpus.

3 Recognizing Entailment

The longest common subsequence (LCS) is used as a measure of similarity between passages. LCS is also used by the ROUGE [10] summarization evaluation package to measure recall of a system summary with respect to a model summary. We use it not to measure recall but precision, to approximate the ratio of information in the hypothesis which is also in the text. Unlike the longest common *substring*, the longest common subsequence does not require adjacency. A longest common subsequence of a text $T = \langle t_1..t_n \rangle$ and a hypothesis $H = \langle h_1..h_n \rangle$ is defined as a longest possible sequence $Q = \langle q_1..q_n \rangle$ with words in Q also being words in T and H in the same order. $\text{LCS}(T, H)$ is the length of the longest common subsequence:

$\text{LCS}(T, H)$

$$= \max \{ |Q| \mid Q \subseteq T \cup H; (t_i = h_k \in Q \wedge t_j = h_l \in Q \wedge j > i) \rightarrow l > k \} \quad (4)$$

From the LCS, the entailment score $\text{LCS}(T, H)/|H|$ is derived. In order to account for variation in natural language text, the LCS is measured after

Table 3. Precision, recall and F-measure for the baseline (100% YES), LCS, LCS after paraphrasing, and dependency tree alignment

lang	baseline			LCS			LCS after paraphr.			tree alignment		
	Pr.	Rec.	F	Pr.	Rec.	F	Pr.	Rec.	F	Pr.	Rec.	F
DE	.251	1.00	.401	.400 ¹	.085 ¹	.140 ¹	.403	.229	.292			
EN	.158	1.00	.273	.312	.181	.229	.304 ²	.479 ²	.372 ²	.343	.512	.410
ES	.294	1.00	.454	.626	.262	.370	.504 ²	.580 ²	.539 ²	.481 ¹	.456 ¹	.468 ¹
FR	.230	1.00	.374	.463 ¹	.052 ¹	.094 ¹	.495	.210	.295			
IT	.172	1.00	.294	.328 ¹	.112 ¹	.167 ¹	.380	.305	.338			
NL	.104	1.00	.188	.217	.160	.184	.199 ²	.346 ²	.252 ²	.287 ¹	.593 ¹	.387 ¹
PT	.237	1.00	.383	.578 ¹	.255 ¹	.354 ¹	.417	.468	.441			

paraphrasing the hypothesis. The underlying idea is that whenever a paraphrase of the hypothesis exists which entails the text, the hypothesis itself also entails the text.

We attempted to extract paraphrases for every phrase in the hypothesis of up to 8 words. Note that by “phrase” we simply mean an (ordered) sequence of words. After generating these candidate mappings we iteratively transform the hypothesis to be closer to the text by substituting in paraphrases. At each iteration, the substitution is made which constitutes the greatest increase of the entailment score. To prevent overgeneration, a word which was introduced in the hypothesis by a paraphrase substitution cannot be substituted itself. The process stops when no more substitutions can be made which positively affect the entailment score. By example, the following paraphrase of the hypothesis from section 1 is obtained by a number of substitutions.

Hypothesis: *Clonaid announced that mother and daughter would be returning to the US on Monday.*

Substitutions:

- the US → the United States
- returning → return
- said → announced
- would be → is
- on Monday → Monday

Paraphrased hypothesis: *Clonaid said that mother and daughter is return to the United States Monday.*

In this case, paraphrasing caused the length of the LCS to increase from 43% ($\frac{6}{14}$) to 77% ($\frac{10}{13}$). The words in italics are the words which are aligned with the text sentence, i.e. which are part of the longest common subsequence. Table 2 shows a number of CLEF AVE pairs for which paraphrases were used to recognize entailment.

¹ These runs are submitted.
² The submitted runs had a slightly higher precision and lower accuracy, due to a preprocessing error.

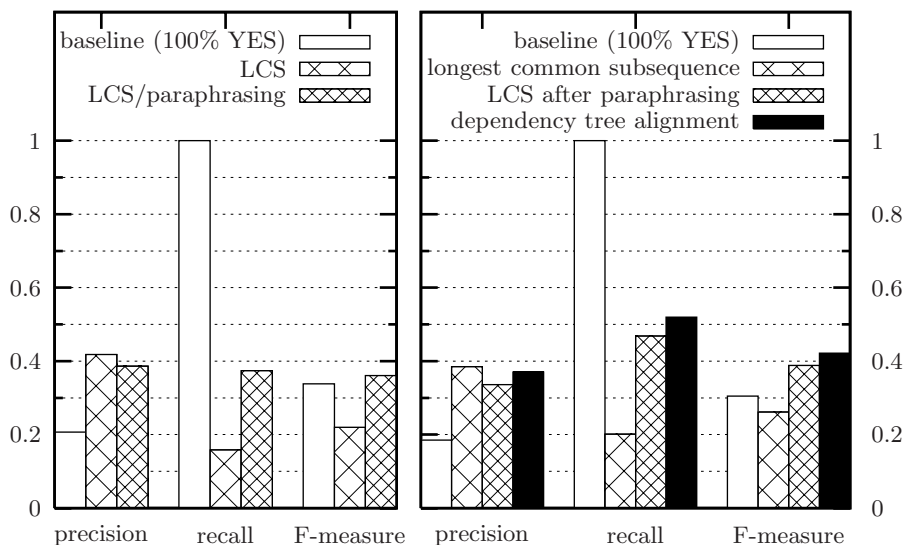


Fig. 2. *Left*, average scores of various metrics for German, English, Spanish, French, Italian, Dutch and Portuguese; *right*, average scores, including dependency tree alignment, for English, Spanish and Dutch

In order to judge whether a hypothesis is entailed by a text we see if the value of the entailment score, $\text{LCS}(T, H)/|H|$, is greater than some threshold value. Support vector machines [11] are used to determine the entailment threshold. Unfortunately, the only suitable training data available was the Question Answering subset of the RTE2 data set [2]. This is a monolingual collection of English passage pairs, with for each pair a boolean annotation of the presence of an entailment relation. Lacking training data for other languages, for our submission we used the RTE2 data to learn the entailment threshold for all languages. The threshold value used throughout these experiments was 0.75.

4 Results

We compared the performance of the paraphrasing method with two baselines on seven languages within the the CLEF 2006 Answer Validation Exercise. The first baseline is a system which always decides that the hypothesis is entailed. The second baseline is a system which measures the longest common subsequence of text and hypothesis. Table 3 lists the performance of the baselines, the paraphrase-based system and the system which uses dependency trees. Average performance over a number of languages is visualized in Figure 3. Given the fact that paraphrasing is a form of query expansion, we expected that precision drops and recall increases when using paraphrases. Results show that this is indeed the case, but that the system using paraphrases shows considerably

Table 4. Entailment assessments by both systems. Percentages of pairs on which the algorithms disagreed are boldfaced. The percentages are averages over Dutch, Spanish and English, compensated for the number of pairs in each language.

LCS after paraphrasing	dependency tree alignment	entailment relation	
		yes	no
yes	yes	5.9%	8.6%
yes	no	3.5%	7.6%
no	yes	3.3%	6.5%
no	no	5.8%	58.8%

better overall performance, as indicated by the F-measure, compared to plain LCS.

For Dutch, Spanish and English, we made a syntactic analysis of each sentence using the parsers of [12], [13] and [14] respectively. As a fourth entailment recognition system, we measured the largest common subtree of the dependency trees of the text and hypothesis. The algorithm of Marsi et al. [15] was used to align dependency trees. Interestingly, as shown in Figure 3 (right), the dependency tree alignment system performs comparably to the largest common subsequence after paraphrasing, while the first uses syntactic information and the latter uses paraphrase generation. The fact that both systems disagree on 37 percent of all pairs with positive entailment (see Table 4) indicates that performance can be further increased when employing both types of information in an integrated system.

5 Conclusion

We evaluated the effect of paraphrasing on a longest common subsequence-based system for recognizing textual entailment. In our CLEF experiments on 7 languages, the system using paraphrases outperformed the system relying on merely the longest common subsequence. Our method is applicable to a wide range of languages, since no language specific natural language analysis or background knowledge is used other than paraphrases automatically extracted from bilingual parallel corpora. Although our system performs similarly to a syntax based system, we showed that there is relatively little overlap between the sets of correctly recognized pairs of both systems. This indicates that information conveyed by paraphrases and syntax are largely complementary for the task of recognizing entailment. In the future we plan to investigate if our system can be improved by using a combination of syntax-based and paraphrase-based approaches to entailment recognition. Also, we plan to improve methods for determining the entailment threshold.

Acknowledgments. This work is funded by the Interactive Multimodal Information Extraction (IMIX) program of the Netherlands Organization for Scientific Research (NWO).

References

1. Vanderwende, L., Dolan, W.B.: What syntax can contribute in the entailment task. In: Quinonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) *MLCW 2005. LNCS (LNAI)*, vol. 3944, pp. 205–216. Springer, Heidelberg (2006)
2. Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The second PASCAL Recognising Textual Entailment Challenge. In: Magnini, B., Dagan, I. (eds.) *Proceedings of the Second PASCAL Recognising Textual Entailment Challenge*, Trento, Italy (April 2006)
3. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the answer validation exercise 2006. In: *Evaluation of Multilingual and Multi-modal Information Retrieval — Seventh CLEF Workshop*, Alicante, Spain. LNCS (September 2006)
4. Barzilay, R., McKeown, K.: Extracting paraphrases from a parallel corpus. In: *ACL-2001* (2001)
5. Pang, B., Knight, K., Marcu, D.: Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In: *Proceedings of HLT/NAACL* (2003)
6. Barzilay, R., Lee, L.: Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In: *Proceedings of HLT/NAACL* (2003)
7. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: *Proceedings of ACL* (2005)
8. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: *Proceedings of HLT/NAACL* (2003)
9. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: *Proceedings of MT-Summit* (2005)
10. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: *Proceedings of ACL 2004 Workshop Text Summarization Branches Out*, Barcelona, Spain (2004)
11. Vapnik, V.N.: *The nature of statistical learning theory*, 2nd edn. Springer, Heidelberg (1999)
12. Bouma, G., van Noord, G., Malouf, R.: Alpino: wide-coverage computational analysis of Dutch. In: *Proceedings of CLIN* (2000)
13. Carreras, X., Chao, I., Padró, L., Padró, M.: FreeLing: an open-source suite of language analyzers. In: *Proceedings of the 4th international Language Resources and Evaluation Conference*, Lisbon, Portugal (2004)
14. Lin, D.: Dependency-based evaluation of MiniPar. In: *Proceedings of LREC Workshop on the Evaluation of Parsing Systems*, Granada, Spain (1998)
15. Marsi, E., Krahmer, E., Bosma, W., Theune, M.: Normalized alignment of dependency trees for detecting textual entailment. In: Magnini, B., Dagan, I. (eds.) *Second PASCAL Recognising Textual Entailment Challenge*, Venice, Italy, pp. 56–61 (2006)

Experimenting a “General Purpose” Textual Entailment Learner in AVE

Fabio Massimo Zanzotto and Alessandro Moschitti

ART Group, DISP, University of Rome “Tor Vergata”, Rome, Italy
{zanzotto,moschitti}@info.uniroma2.it
<http://ai-nlp.info.uniroma2.it>

Abstract. In this paper we present the use of a “general purpose” textual entailment recognizer in the Answer Validation Exercise (AVE) task. Our system is designed to learn entailment rules from annotated examples. Its main feature is the use of Support Vector Machines (SVMs) with kernel functions based on cross-pair similarity between entailment pairs. We experimented with our system using different training sets: RTE and AVE data sets. The comparative results show that entailment rules can be learned. Although, the high variability of the outcome prevents us to derive definitive conclusions, the results show that our approach is quite promising and improvable in the future.

1 Introduction

Textual entailment recognition is a common task performed in several natural language processing applications [1], e.g. Question Answering and Information Extraction. The Recognizing Textual Entailment (RTE) PASCAL Challenges [2,3] fostered the development of several “general purpose” textual entailment recognizers. The Answer Validation Exercise (AVE) instead provides an opportunity to show that such recognizers are useful for Question Answering.

We applied our textual entailment recognition system [4], developed for the second RTE challenge [3], to AVE. Our system has been shown to achieve state-of-the-art results on both RTE data sets [2,3]. It determines whether or not a text T entails a hypothesis H by automatically learning rewriting rules from positive and negative training instances of entailment pairs (T, H) . For example given a text T_1 : “*At the end of the year, all solid companies pay dividends.*” and two hypotheses:

- a) H_1 : “*At the end of the year, all solid insurance companies pay dividends*” and
- b) H_2 : “*At the end of the year, all solid companies pay cash dividends*”,

we can build two examples: (T_1, H_1) which is a true entailment (positive instance) and (T_1, H_2) which is a negative one. Our system extract rules from them to solve apparently not related entailments. For example, given the following text and hypothesis:

$$T_3 \Rightarrow H_3?$$

T_3	“All wild animals eat plants that have scientifically proven medicinal properties.”
H_3	“All wild mountain animals eat plants that have scientifically proven medicinal properties.”

we note that T_3 is structurally (and somehow lexically similar) to T_1 and H_3 is more similar to H_1 than to H_2 . Thus, from $T_1 \Rightarrow H_1$, we may extract rules to derive that $T_3 \Rightarrow H_3$.

The main idea of our model is that it relies not only on a *intra-pair* similarity between T and H but also on a *cross-pair* similarity between two pairs (T', H') and (T'', H'') . The latter similarity measure along with a set of annotated examples allows the leaning model to automatically derive syntactic and lexical rules that can solve complex entailment cases.

In this paper, we experimented with our textual entailment recognition system [4] and the CLEF AVE. We show that entailment recognition rules can be learnt from annotated examples. In the remainder of this paper, Sec. 2 describes our model in a glance, Sec. 3 shows the experimental results, and, finally, Sec. 4 derives the conclusions.

2 Our Textual Entailment Learner in a Glance

In this section we will shortly introduce the main idea of our method, which is fully described in [4].

To carry out automatic learning from examples, we need to define a cross-pair similarity $K((T', H'), (T'', H''))$. This function should consider pairs similar when: (1) texts and hypotheses are structurally and lexically similar (*structural similarity*); (2) the relations between the sentences in the pair (T', H') are compatible with the relations in (T'', H'') (*intra-pair word movement compatibility*). We argue that such requirements could be met by augmenting syntactic trees with *placeholders* that co-index related words within pairs. We will then define a cross-pair similarity over these pairs of co-indexed trees.

2.1 Training Examples as Pairs of Co-indexed Trees

Sentence pairs selected as possible sentences in entailment are naturally co-indexed. Many words (or expressions) w_h in H have a referent w_t in T . The pairs (w_t, w_h) are called *anchors*, where different relations between w_t and w_h may be hold. Therefore, entailment could hold even if the two words are substituted with two other related words. To indicate the relatedness property, we co-index words with *placeholders* which in turn are associated with anchors. For example, in Fig. 1, [2] indicates the *(companies, companies)* anchor between T_1 and H_1 . These placeholders are then used to augment tree nodes. To better take into account argument movements, placeholders are propagated in the syntactic trees following constituent heads (see Fig. 1).

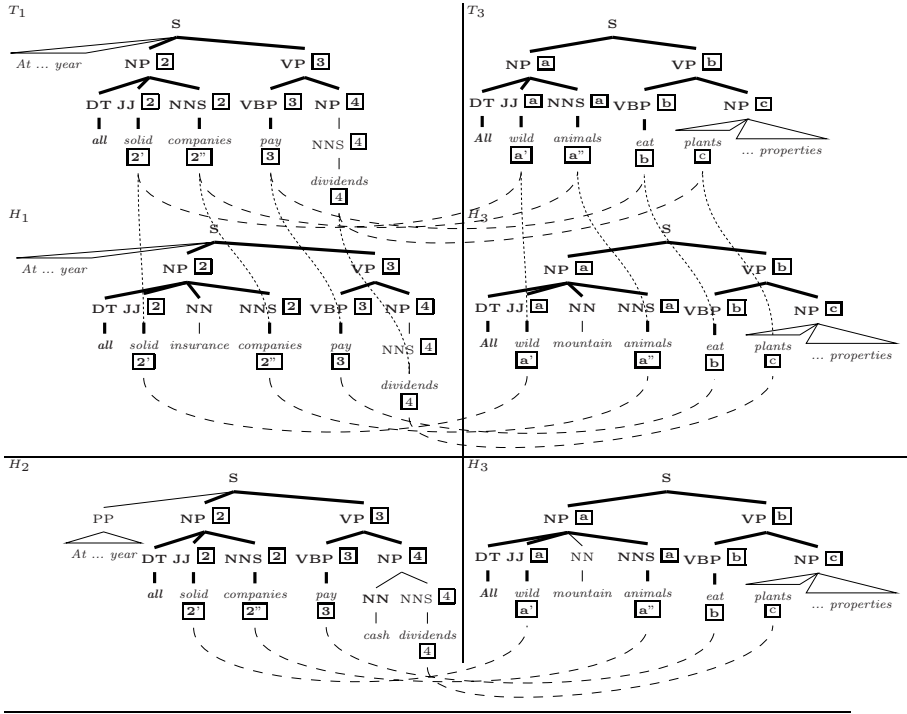


Fig. 1. Relations between (T_1, H_1) , (T_1, H_2) , and (T_3, H_3)

In line with many other researches (e.g., [5]), we determine these anchors using different similarity or relatedness models: the exact matching between tokens or lemmas, a similarity between tokens based on their edit distance, the derivationally related form relation and the verb entailment relation in WordNet, and, finally, a WordNet-based similarity [6]. Each of these detectors gives a different weight to the anchor: 0/1 for the first and the similarity value for all the others. These weights will be used in the final kernel.

2.2 Similarity Between Pairs of Co-indexed Trees

Pairs of syntactic trees where nodes are co-indexed with placeholders allow us to design a cross-pair similarity based on both the structural similarity and the intra-pair word movement compatibility.

Syntactic trees of texts and hypotheses are useful to detect structural similarity between pairs of sentences. Texts should have similar structures as well as hypotheses. In Fig. 1, the overlapping subtrees are in bold. For example, T_1 and T_3 share the subtree starting with $S \rightarrow NP VP$. Although the lexicals in T_3 and H_3 are quite different from those T_1 and H_1 , their bold subtrees are more

similar to those of T_1 and H_1 than to T_1 and H_2 , respectively. H_1 and H_3 share the production $\mathbf{NP} \rightarrow \mathbf{DT JJ NN NNS}$ while H_2 and H_3 do not. To decide on the entailment for (T_3, H_3) , we can use the decision made on (T_1, H_1) .

Anchors and placeholders are useful to verify if two pairs can be aligned by checking the compatibility between intra-pair word movements. For example, (T_1, H_1) and (T_3, H_3) show compatible constituent movements given that the dashed lines connecting placeholders of the two pairs indicate structurally equivalent nodes both in the texts and the hypotheses. The dashed line between $\boxed{3}$ and $\boxed{b}$ links the main verbs both in the texts T_1 and T_3 and in the hypotheses H_1 and H_3 . After substituting $\boxed{3}$ to $\boxed{b}$ and $\boxed{2}$ to $\boxed{a}$, T_1 and T_3 share the subtree $\mathbf{S} \rightarrow \mathbf{NP}\boxed{2} \mathbf{VP}\boxed{3}$. The same subtree is shared between H_1 and H_3 . This implies that words in the pair (T_1, H_1) are correlated like words in (T_3, H_3) . Any different mapping between the two anchor sets would not have this property.

Using the structural similarity, the placeholders, and the connection between placeholders, the overall similarity $K_s((T', H'), (T'', H''))$ is then defined on a kernel $K_T(st_1, st_2)$ (as the one described in [7]) that measures the similarity between two syntactic trees st_1 and st_2 . For more details on the kernel function see [4].

3 Experimental Investigation

The experiments aim at determining if our system can learn rules required to solve the entailment cases contained in the AVE data set. Although, we have already shown that our system can learn entailment rules [2,3], the task here appears to be more complex as: (a) texts are automatically built from answers and questions; this necessarily introduces some degree of noise; and (b) often question answering systems provide a correct answer but the supporting text is not adequate to carry out a correctness inference, e.g. a lot background knowledge is required or the answer was selected by chance.

Our approach to study the above points is to train and experiment with our system and several data sets proposed in AVE as well as in the RTE challenges. The different training based on such sets can give an indication on the learnability of general rules valid for different domain and different applications.

3.1 Experimented Kernels

We compared three different kernels: (1) the kernel $K_l((T', H'), (T'', H'')) = sim_l(T', H') \times sim_l(T'', H'')$ which is based on the intra-pair lexical similarity $sim_l(T, H)$ described in [5]. (2) The kernel $K_l + K_s$ that combines our kernel with the lexical-similarity-based kernel. (3) The kernel $K_l + K_t$ that combines the lexical-similarity-based kernel with a basic tree kernel. The latter is defined as $K_t((T', H'), (T'', H'')) = K_T(T', T'') + K_T(H', H'')$.

3.2 Experimental Settings

For the experiments, we used the following data sets:

- *RTE*, i.e. the set derived using all the data (development and test data) of the first [2] and second [3] RTE challenges. Respectively, the dataset of the first RTE contains 1,367 examples whereas the one of the second contains 1,600 instances. The positive and negative examples are equally distributed in the collection, i.e. 50% of the data.
- *AVE*, i.e., the AVE development set. The AVE development set contains 2870 instances. Here, the positive and negative examples are not equally distributed. It contains 436 positive and 2434 negative examples.
- *AVE_{test}*, i.e., the AVE testing set. The AVE test set contains 2088 instances. It contains 198 positive 1048 and negative examples. The remaining 842 examples are unknown (see the AVE task overall description [8] for more details) and are not used for evaluation purposes.

We also created a new set $AVE \cup RTE$ by merging groups of the two of the above collections.

In the remainder of this section, we describe the resources used to implement our system.

- The Charniak parser [9] and the **morpha** lemmatiser [10] to carry out the syntactic and morphological analysis.
- WordNet 2.0 [11] to extract both the verbs in entailment, *Ent* set, and the derivationally related words, *Der* set.
- The **wn::similarity** package [12] to compute the Jiang&Conrath (J&C) distance [6] as in [5]. This is one of the best figure method which provides a similarity score in the $[0, 1]$ interval. We used it to implement the $d(l_w, l_{w'})$ function.
- A selected portion of the British National Corpus¹ to compute the inverse document frequency (*idf*). We assigned the maximum *idf* to words not found in the BNC.
- SVM-light-TK² [13] which encodes the basic tree kernel function, K_T , in SVM-light [14]. We used such software to implement the kernels K_l , K_s , and K_t and their linear combinations.

3.3 Results and Analysis

Table 1 reports the results of our system trained with data from RTE and AVE development and tested on the AVE test set (*AVE_{test}*). Column 1 denotes the data set used for training. Column 2 illustrates the value of the j parameter, respectively. Such parameter tunes the trade-off between Precision and Recall. Higher values cause the system to retrieve more positive examples. We used

¹ <http://www.natcorp.ox.ac.uk/>

² SVM-light-TK is available at <http://ai-nlp.info.uniroma2.it/moschitti/>

Table 1. The results of the different models over the test set AVE_{test}

Model		K_l			$K_s + K_l$			$K_t + K_l$		
Train	j	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
AVE	10	0.2512	0.8232	0.3849	0.3858	0.4949	0.4336	0.3979	0.3838	0.3907
RTE	0.9	0.2916	0.8232	0.4307	0.2944	0.7121	0.4166	0.2872	0.7121	0.4093
	1	0.2820	0.8232	0.4201	0.3008	0.7778	0.4338	0.2729	0.7525	0.4005
	10	0.1589	1.0000	0.2742	0.2131	0.9343	0.3470	0.2305	0.9697	0.3725
$RTE \cup AVE$	0.9	0	0.0000	0	0.3431	0.4747	0.3983	0.2971	0.5101	0.3755
	1	0	0.0000	0	0.3399	0.6061	0.4355	0.2844	0.6364	0.3931
	10	0.1589	1.0000	0.2742	0.2290	0.8384	0.3597	0.2408	0.8939	0.3794

three values: 0.9, 1, and 10. Row 1 indicates the kernel that has been used for the experiments. The Precision, the Recall, and the F-measure are computed on the positive pairs (yes pairs) as discussed in the AVE task description [8]. When the system has a Recall of 0, we assign Precision and the F-Measure to 0. Results for the systems trained on the AVE with j equals to 0.9 and 1 are not in the table as they all show a 0 Recall. The revised³ results of the two runs submitted are in bold. ZNZ-TV.1 corresponds to the $K_l + K_s$ system trained on the AVE set with $j = 10$ whereas ZNZ-TV.2 corresponds to the $K_l + K_s$ system trained on the RTE set with $j = 0.9$.

The following aspects should be noted:

- When learning on the AVE set, the small amount of positive examples limits the possibility of learning the positive entailments. The resulting model classifies always as negative example. The only way of extracting positive rules is to give more importance to positive examples changing the j trade-off parameter. Only with the value 10, a model retrieving positive examples is learnt.
- The model $K_l + K_s$ is generally better than the model $K_l + K_t$. For example, for the submitted system trained on AVE, the difference is of about 5 points in term of F-Measure. This indicates that placeholders are very relevant for learning entailment rules.
- The simple model K_l learnt on the RTE has an F-Measure comparable with those obtained by the best model, i.e. 0.4336 of $K_l + K_s$ system trained on the AVE set. This odd result is generated by the high variability of system accuracy, e.g. slightly different parameterization produce very different results.

Qualitative analysis. The system we presented strongly uses syntactic interpretations of the example pairs. Then, its major bottleneck is the standard AVE process used to produce the affirmative form of the question given the answer provided by a the QA system. This process frequently generates ungrammatical sentences. The problem is clear just reading the first instances of the AVE development set. We report hereafter some of these examples. Each table reports

³ Results presented here differ from the official results. At the time of the submission, the K_s part of the combined kernel $K_l + K_s$ didn’t work properly for a formatting error of the syntactic trees. Basically, the systems submitted were the systems based on K_l with some noise.

the original question (Q), the text snippet (T), and the affirmative form of the question used as hypothesis (H).

$T \Rightarrow H$ (id: 2)	$T \Rightarrow H$ (id: 6)
Q “What year was Halley’s comet visible?”	Q “Who is Juan Antonio Samaranch?”
T “[...] 1909 Halley’s comet sighted from Cambridge Observatory. 1929 [...]”	T “International Olympic Committee President Juan Antonio Samaranch came strongly to the defense of China’s athletes, [...]”
H “In 1909 Halley was Halley’s comet visible ”	H “Juan Antonio Samaranch is International Olympic Committee President Juan Antonio Samaranch came strongly to the defense of China ’s athletes ”

We can observe that these examples have highly ungrammatical hypotheses. In the example (id 2), *Halley* is used as subject and as predicate. Finally, in the example (id 6) there is a large part of the hypothesis that is unnecessary and creates an ungrammatical sentence.

4 Conclusions

In this paper, we experimented with our entailment system [4] and the CLEF AVE. The comparative results show that entailment rules can be learned from the data set AVE. The experiments show that few training examples and data sparseness produce a high variability of the results. In this scenario the parameterization is very critical and necessitates of accurate cross-validation techniques. In the future, we would like to carry out a throughout parameterization and continue investigating approaches to exploit data from difference sources of entailment rules.

References

1. Dagan, I., Glickman, O.: Probabilistic textual entailment: Generic applied modeling of language variability. In: Proceedings of the Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France (2004)
2. Dagan, I., Glickman, O., Magnini, B.: The PASCAL RTE challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d’Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, Springer, Heidelberg (2006)
3. Bar Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I.: The II PASCAL RTE challenge. In: PASCAL Challenges Workshop, Venice, Italy (2006)
4. Zanzotto, F.M., Moschitti, A.: Automatic learning of textual entailments with cross-pair similarities. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia. Association for Computational Linguistics, pp. 401–408 (July 2006)

5. Corley, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proc. of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, Michigan. Association for Computational Linguistics, June 13–18, 2005 (2005)
6. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the 10th ROCLING, Tapei, Taiwan, pp. 132–139 (1997)
7. Collins, M., Duffy, N.: New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In: Proceedings of ACL02 (2002)
8. Peñas, A., Rodrigo, A., Sama, V., Verdejo, F.: Overview of the answer validation exercise 2006. In: Peters, C., Clough, P., Gey, F., Karlgren, J., Magnini, B., Oard, D., de Rijke, M., Stempfhuber, M. (eds.) Evaluation of Multilingual and Multimodal Information Retrieval. LNCS, Springer, Heidelberg (2007)
9. Charniak, E.: A maximum-entropy-inspired parser. In: Proc. of the 1st NAACL, Seattle, Washington, pp. 132–139 (2000)
10. Minnen, G., Carroll, J., Pearce, D.: Applied morphological processing of English. *Natural Language Engineering* 7(3), 207–223 (2001)
11. Miller, G.A.: WordNet: A lexical database for English. *Communications of the ACM* 38(11), 39–41 (1995)
12. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet: similarity - measuring the relatedness of concepts. In: Proc. of 5th NAACL, Boston (2004)
13. Moschitti, A.: Making tree kernels practical for natural language learning. In: Proceedings of EACL’06, Trento, Italy (2006)
14. Joachims, T.: Making large-scale svm learning practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) *Advances in Kernel Methods-Support Vector Learning*, 1999, Cambridge (1999)

Answer Validation Through Robust Logical Inference

Ingo Glöckner

Intelligent Information and Communication Systems (IICS)
University of Hagen, 58084 Hagen, Germany
iglockner@web.de

Abstract. The paper features MAVE, a knowledge-based system for answer validation through deep linguistic processing and logical inference. A relaxation loop is used to determine a robust indicator of logical entailment. The system not only validates answers directly, but also gathers evidence by proving the original question and comparing results with the answer candidate. This method boosts recall by up to 13%. Additional indicators signal false positives. The filter increases precision by up to 14.5% without compromising recall of the system.

1 Introduction

MAVE (MultiNet-based answer verification), a system for validating results of question-answering (QA) systems, was developed as a testbed for robust knowledge processing in the MultiNet paradigm [1]. The goal was that of showing that deep NLP and logical inference (plus relaxation) are robust enough to handle the validation task successfully.

2 System Description

2.1 Overview of MAVE

A validation candidate p submitted to MAVE consists of a hypothesis string constructed by substituting the answer of the QA system into the original question, the snippet with the relevant text passage and also of the original question. Processing of input p involves these steps: a) *Preprocessing*, i.e. application of correction rules based on regular expressions. This correction step is necessary since the mechanism for generating hypotheses often results in syntactically ill-formed hypotheses. Typical errors like ‘*in in Lillehammer*’ are easily eliminated, however. b) The preprocessed inputs are subjected to a *deep linguistic analysis* by the WOCADI parser [2] which results in a MultiNet representation [1]. This process includes disambiguation of word meanings, semantic role labeling, a facticity labeling, and coreference resolution. c) The post-processing involves a *normalization* of the generated semantic networks. A list of about 1,000 synonyms from the computational lexicon HaGenLex [3] and additional 525 hand-crafted synonyms are used for replacing lexical concepts with a canonical synset representative. Not only the query and snippet, but also the known facts and implicative rules are automatically normalized in this way. d) Next is *query construction*, i.e. translation of the generated semantic network into a conjunction of literals. Notice that two logical

queries are constructed one for the hypothesis and one for the original question. e) *Robust entailment proving* is then tried for the query constructed from the hypothesis and the query constructed from the question (see next section). f) Additional non-logical indicators are extracted, which help detect trivial answers, wrong analyses of the NLP components and other sources of error. g) The results of the logical and non-logical indicators are aggregated into a validity score from which a YES/NO decision is obtained by applying a threshold.

2.2 Robust Knowledge Processing in MAVE

Apart from the logical representation of the snippet, the knowledge of MAVE comprises 2,484 basic facts most of which describe lexico-semantic relationships. 103 implicative rules define key properties of the MultiNet relations and knowledge on biographic topics like birth of a person. The MAVE prover uses iterative deepening to find answers. Performance is optimized by term indexing and literal sorting under a least-effort heuristics [4]. To achieve robust logical inference, the prover is embedded in a *relaxation loop*. The prover keeps track of the longest prefix $L_1, \dots, L_k$ of the literal list for which a proof was found. When a full proof fails, the literal L_{k+1} will be skipped and a new proof of the smaller fragment begins. By subsequently removing ‘critical’ literals, this process always finds a (possibly empty) query fragment provable from the given knowledge. The number of skipped literals serves as a numeric indicator of entailment robust against slight errors in representations and also against gaps in the knowledge.

2.3 Indicators and Scoring Function

MAVE uses the following validity indicators:

- **hypo-triviality** – *Lemma repetition ratio*. This criterion measures the ratio of lemmata which occur an even number of times in the hypothesis. Trivial hypotheses like ‘*Gianni Versace was Gianni Versace*’ result in a high repetition score.
- **hypo-num-sentences** – *Number of sentences in the hypothesis*. The parser might wrongly consider a hypothesis as consisting of two sentences, which indicates an erroneous NL analysis.
- **hypo-collapse** – *Query extraction control index*. This indicator relates the size of the constructed query to the number of content words in the hypothesis. A small value indicates a failure of semantic analysis which produced an incomplete query.
- **hypo-lexical-focus** and **question-lexical-focus** – *Hint at potential false positive*. This criterion indicates that the logical hypothesis representation (representation of original question) is too permissive due to a systematic parser error.
- **hypo-missing-constraints** and **question-missing-constraints** – *Control index for numerical constraints*. Apart from regular words, the hypothesis (original question) can also contain numbers. These numbers must show up as restrictions in the constructed logical representations (e.g. a restriction of the year to 1994), but are occasionally dropped by the parser, as detected by these indicators.
- **hypo-failed-literals** and **question-failed-literals** – *Number of non-provable literals*. The relaxation loop into which the prover is embedded determines the number of non-provable literals in the hypothesis representation (representation of the original question). A value of 0 means strict logical entailment.

- *hypo-delta-concepts* – *Number of answer words in the hypothesis*. The indicator then counts the number of content-bearing tokens and in the hypothesis which stem from the original answer of the QA system and not from the question.
- *hypo-delta-matched* – *Number of matched answer words*. Based on the proof of the logical question representation over the snippet and the known origin of literals in one of the sentences of the snippet, MAVE identifies that sentence in the snippet which contains most literals supporting the proof. The content-bearing lexemes and numbers in the answer part of the hypothesis are matched against the lexical concepts and numbers in the MultiNet representation of the central sentence.

In Scheme notation, the imperfection score $\text{imp-score}(p)$ of p is given by:

```
(max (* (max (>= hypo-triviality 0.8) (= hypo-num-sentences 2)) 1000)
  (min (max (* (max (< hypo-collapse 0.7) hypo-lexical-focus) 1000)
    (+ hypo-missing-constraints hypo-failed-literals))
    (max (* question-lexical-focus (= hypo-delta-matched 0) 1000)
      (+ question-failed-literals question-missing-constraints
        (- hypo-delta-concepts hypo-delta-matched))))
```

For details of handling undefined values, see [4]. A threshold θ is used to determine YES/NO decisions, i.e. p is accepted if $\text{imp-score}(p) \leq \theta$ and rejected otherwise.

3 Evaluation

Two runs were submitted to AVE 2006. The recall-oriented first run allowed $\theta = 2$ mismatches for YES decisions. The precision-oriented second run allowed $\theta = 1$ mismatch.

Table 1 lists the results of MAVE in the answer validation exercise. ‘Accuracy’ is ratio of correct decision. Although the approach does not search for a minimal set of failed literals, but only follows a ‘longest partial proof’ strategy, the failed literal counts are significant with respect to validity of entailment: There is a decrease in precision if more imperfections θ (usually skipped literals) are allowed, but also a boost in recall. Moreover the two submitted runs represent the best compromise. Assuming perfect provability ($\theta = 0$) loses too much recall, while $\theta \geq 3$ mismatches considerably lower precision. MAVE also resorts to question proofs for validating hypotheses. This mechanism is intended to increase recall when the hypothesis is syntactically ill-formed. The hypothesis is then compared with the answer determined by the proof of the question in a way which does not assume a parse of the hypothesis. Table 2 (left) lists the results when no question-related indicators are used. For $\theta = 1$, recall decreases by 10.2 percent points. For $\theta = 2$, question indicators even contribute 13.1 percent points to the achieved recall. Precision is lowered somewhat by using the additional question data but remains on a high level. The definition of $\text{imp-score}(p)$ involves false-positive indicators (like *hypo-triviality*, *hypo-collapse* or *question-lexical-focus*). Table 2 (right) shows the results of MAVE without these indicators. Precision drops by 14.5 percent points when false positive indicators are ignored, thus proving their effectiveness. Recall is not altered significantly, i.e. the criteria are also sufficiently specific.

Table 1. Results of MAVE for the AVE-2006 test set

θ (run)	precision	recall	f-measure	accuracy
0	0.8584	0.2820	0.4245	0.8132
1 (run #2)	0.7293	0.3837	0.5029	0.8146
2 (run #1)	0.5839	0.5058	0.5421	0.7912
3	0.4786	0.5843	0.5262	0.7429
4	0.4024	0.6831	0.5065	0.6747
5	0.3592	0.7413	0.4839	0.6136

Table 2. Results for hypothesis proofs only (left), omitting false positive tests (right)

θ	precision	recall	f-measure	accuracy	θ	precision	recall	f-measure	accuracy
0	0.9452	0.2006	0.3309	0.8018	0	0.7132	0.2820	0.4042	0.7969
1	0.8362	0.2820	0.4217	0.8111	1	0.6000	0.4012	0.4808	0.7884
2	0.6898	0.3750	0.4859	0.8061	2	0.5043	0.5116	0.5079	0.7578
3	0.5510	0.4709	0.5078	0.7770	3	0.4238	0.5901	0.4933	0.7038
4	0.4450	0.5407	0.4882	0.7230	4	0.3652	0.6890	0.4773	0.6314
5	0.3900	0.6134	0.4768	0.6712	5	0.3329	0.7471	0.4606	0.5724

4 Conclusion

MAVE achieved the best results for German in the AVE 2006 exercise. This success is mainly due to the robust proving strategy which tolerates a few mismatches and gaps in the coded knowledge. More ablation studies on the contribution of individual system components to the achieved filtering quality can be found in [5]. A successful application of MAVE for improving answer quality by validating and merging results in a multi-stream QA setting is described in [6].

References

1. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Heidelberg (2006)
2. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück, Germany (2003)
3. Hartrumpf, S., Helbig, H., Osswald, R.: The semantically based computer lexicon HaGenLex. *Traitement automatique des langues* 44(2), 81–105 (2003)
4. Glöckner, I.: University of Hagen at QA@CLEF 2006: Answer validation exercise. In: *Working Notes for the CLEF06 Workshop*, Alicante, Spain (2006)
5. Glöckner, I.: Filtering and fusion of question-answering streams by robust textual inference. In: *Proc. of KRAQ'07*, Hyderabad/India (2007)
6. Glöckner, I., Hartrumpf, S., Leveling, J.: Logical validation, answer merging and witness selection – A study in multi-stream question answering. In: *Proc. of RIAO-07* (2007)

University of Alicante at QA@CLEF2006: Answer Validation Exercise

Zornitsa Kozareva, Sonia Vázquez, and Andrés Montoyo

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{zkozareva,svazquez,montoyo}@dlsi.ua.es

Abstract. In this paper we present an Answer Validation System which is based on the combination of word overlap and Latent Semantic Indexing modules. The main contribution of our work consist in the adaptation of our already developed machine-learning textual entailment system MLEnt to the multilingual Answer Validation exercise.

1 Introduction

The task of textual entailment recognition (RTE)¹ [1] consists in determining whether the meaning of one text can be inferred from the meaning of another. RTE is related to the Answer Validation Exercise (AVE) [5] which is centered on the validation of the Question Answering (QA) questions and text snippets returned by the QA system. Therefore, we decided to evaluate our already developed RTE system MLEnt [2] in the AVE challenge.

One characteristic of MLEnt's system is that it does not use any language dependent tools or resources, which means that the system can be easily adapted to different languages. However, we could not verify our claims in the RTE challenge since there is no multilingual textual entailment corpora. Due to the relatedness of the RTE and AVE tasks, and the fact that the AVE organizers provide multilingual data, we decided to evaluate the performance of our approach and verify the MLEnt's hypothesis of multilinguality. For this reason, we have evaluated the system with the English, Spanish, French, Dutch, German, Italian and Portuguese languages, and also we have studied the effect of lexical and semantic classifier combination on the basis of voting strategies.

2 System Description

The AVE system [3] consists of two modules: a word overlap and Latent Semantic Indexing ones which are combined through voting.

¹ <http://www.pascal-network.org/Challenges/RTE/>
<http://www.pascal-network.org/Challenges/RTE2/>

2.1 Word Overlap Module

The first word overlap attribute is *n-gram*. It looks for common position-independent unigram matches between a question Q and a text snippet T in the AV pair. According to this attribute, the AV pair is correct when Q and T have many common words, and the AV pair is incorrect when there are no common words at all.

The second attribute is *LCS*. It estimates the similarity between Q with length m and T with length n , as $\frac{LCS(Q,T)}{n}$ identifying common non-consecutive word sequences of any length. Longer *LCS* correspond to more similar sentences and correct AV pair.

The third attribute is *skip-gram*. It represents any pair of words in sentence order that allows arbitrary gaps. Once all word pairs for Q and T are generated, the overlapping *skip-grams* are counted using the measure $skip\_gram = \frac{skip\_gram(Q,T)}{C(n, number\_of\_skip\_gram)}$, where $skip\_gram(Q,T)$ refers to the number of common *skip-grams* found in Q and T, $C(n, number\_of\_skip\_gram)$ is a combinatorial function, where n is the number of words in T and $number\_of\_skip\_grams$ corresponds to the number of common n-grams between (Q,T)². AV pair is correct when Q and T have more common *skip-grams*.

The final word overlap attribute is the *number matching* which identifies the numeric expressions in Q and T, and verifies whether they contain the same numeric values or not. According to this attribute, the AV pair is correct when the numbers in Q and T coincide.

2.2 Latent Semantic Indexing Module

In order to build the conceptual space of LSI [4], we used the T phrases as documents and the Q phrases as terms. Each Q-T matrix is reduced to 300 dimensions using Singular Value Decomposition algorithm. At the end of this process, LSI returns a score between 0 and 1 which indicates the semantic similarity relatedness between the question and the text snippet. Experimentally, we have evaluated that a similarity threshold $\tau > 0.8$ correspond to correct Q-T pair.

In the final stage of our AVE system, the previously described modules are combined through majority voting. According to the experimental results, the best combinations are between *LCS* with *skip-gram*, and *LCS*, *skip-gram*, LSI.

3 Results and Discussion of the Multilingual AVE Runs

Table 1 shows the results from the multilingual AVE. For the English and Spanish runs, we have trained the AVE system with the ENGARTE and SPARTE data sets³. These results are used as an indicator for the attribute selection process

² (e.g. *number_of_skip-grams* is 1 if there is a common unigram between Q and T, 2 if there is a common bigram etc.).

³ <http://nlp.uned.es/QA/ave>

for the rest of the languages for which there is no training data. For English, the best features for the train and test data sets are obtained with the *LCS* and *skip-grams* attributes. However, the highest score is obtained when *LCS* and *skip-grams* are combined with the semantic information provided by LSI. This shows that LSI resolves correctly AV examples which are omitted by the lexical module, and that the voting combination improves the individual performance of the classifiers with 8%. For Spanish, the best performance is yielded with the *LCS* attribute and the voting combination of *LCS*, *skip-gram* and LSI. For German, French and Italian, the best performances are also obtained with the *LCS* feature. The voting runs range from 40 to 47%. The lowest scores are obtained for Dutch, due to the structure of the language which affects the *skip-gram* and *LCS* word-overlapping attributes. The voting combination for Dutch improved the performance with 4%.

Table 1. Results for the AVE runs

Language runs	Precision	Recall	F-score	Language runs	Precision	Recall	F-score
<i>ENGh_LCS</i>	15.22	80.93	28.57	<i>FR_LCS</i>	33.96	67.09	45.09
<i>ENG_Skip</i>	16.91	69.30	27.18	<i>FR_Skip</i>	30.48	46.38	36.78
<i>ENG_LSI</i>	23.29	30.23	26.31	<i>FR_LSI</i>	32.36	15.88	21.31
<i>ENG_LCS&Skip</i>	18.33	64.65	28.56	<i>FR_LCS&Skip</i>	38.36	43.69	40.85
<i>ENG_LCS&Skip&LSI</i>	24.92	69.77	36.72	<i>FR_LCS&Skip&LSI</i>	34.44	73.62	46.93
<i>SP_LCS</i>	44.21	66.62	53.15	<i>IT_LCS</i>	25.78	70.59	37.77
<i>SP_Skip</i>	37.24	43.07	39.94	<i>IT_Skip</i>	21.96	86.10	34.99
<i>SP_LSI</i>	34.15	14.45	20.31	<i>IT_LSI</i>	29.16	22.45	25.37
<i>SP_LCS&Skip</i>	47.48	39.34	43.03	<i>IT_LCS&Skip</i>	21.64	88.77	34.80
<i>SP_LCS&Skip&LSI</i>	40.65	76.15	53.01	<i>IT_LCS&Skip&LSI</i>	28.30	72.19	40.66
<i>GER_LCS</i>	38.90	60.56	47.37	<i>DU_LCS</i>	14.26	90.12	24.62
<i>GER_Skip</i>	34.37	43.91	38.55	<i>DU_Skip</i>	15.80	67.901	25.64
<i>GER_LSI</i>	11.43	1.13	2.06	<i>DU_LSI</i>	13.88	12.34	13.07
<i>GER_LCS&Skip</i>	41.30	37.68	39.41	<i>DU_LCS&Skip</i>	18.90	67.90	29.57
<i>GER_LCS&Skip&LSI</i>	36.34	67.42	47.22	<i>DU_LCS&Skip&LSI</i>	14.84	90.12	25.48
<i>PT_LCS</i>	12.50	3.90	5.94	<i>PT_Skip</i>	8.00	21.00	11.58
<i>PT_LSI</i>	11.26	12.76	11.96	<i>PT_LCS&Skip</i>	19.04	12.77	15.29

Globally for the Spanish, German and French languages, the *LCS* and voting combinations performed the same. For English, the voting combination improved the performance with 8%, while for Italian and Portuguese there was an increment of 3%. Compared to the individual performance of LSI, the voting combination performed better for each one of the seven languages.

4 Conclusions and Future Work

The main aim for our participation in the AVE competition was to test our previously developed textual entailment system MLEnt for multilinguality and how it will perform on a new task such as AV. As the textual entailment and the answer validation tasks share the same idea of identifying whether two texts infer the same meaning or not, it was easy and possible to adapt the MLEnt system.

With the obtained results, we prove that MLEnt can function for different languages. The performance of the system ranges from 20% to 53% depending on the language and the amount of data. It is interesting to note that the

performance of our system depends on a set of lexical and semantic attributes, however the two most robust attributes are *LCS* and *skip-grams* which for German, French, Italian and Dutch reached 47% and 38% respectively.

During the experiments, the voting combination lead to improvement for the English, French, Italian and Portuguese languages. For Spanish and German the performance of voting and the performance of the *LCS* attribute differed by 0.14% which according to the z' statistics is insignificant. For Dutch the *LCS* and *skip-gram* attributes performed better than voting. According to the conducted experiments, the most informative attribute for the correctness of the AV pair is *LCS*.

The current AVE system has the ability to work fast and is characterized by quick training and testing phases, which is very important for a real Question Answering application. Another benefit for our AVE system is coming from the nature of the attributes, which depend only on the length of the sentences, the number of overlapping words and the context which allows a multilingual resolution.

In the future, we want to measure the similarity between noun phrases and to validate the named entities in the question and the text snippet.

Acknowledgements

This research has been partially funded by the European Union under the project QALLME number FP6 IST-033860 and by the Spanish Ministry of Science and Technology under the project TEX-MESS number TIN2006-15265-C06-01.

References

1. Dagan, I., Glickman, O.: Probabilistic textual entailment: Generic applied modeling of language variability. In: Proceedings of the PASCAL workshop on Text Understanding and Mining
2. Kozareva, Z., Montoyo, A.: Mlent: The machine learning entailment system of the university of alicante. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (2006)
3. Kozareva, Z., Vazquez, S., Montoyo, A.: Adaptation of a machine-learning textual entailment system to a multilingual answer validation exercise. In: Working Notes of CLEF 2006 (2006) ISBN 2-912335-21-3
4. Kozareva, Z., Vázquez, S., Montoyo, A.: The effect of semantic knowledge expansion to textual entailment recognition. In: TSD, pp. 143–150 (2006)
5. Peñas, A., Rodrigo, A., Verdejo, F.: Sparte, a test suite for recognising textual entailment in spanish. In: Proceedings of CICLing (2006)

Towards Entailment-Based Question Answering: ITC-irst at CLEF 2006

Milen Kouylekov, Matteo Negri, Bernardo Magnini, and Bonaventura Coppola

Centro per la Ricerca Scientifica e Tecnologica ITC-irst,
Via Sommarive, 18 Povo-Trento, Italy
`{kouylekov,magnini,negri,coppolab}@itc.it`

Abstract. This year, besides providing support to other groups participating in cross-language Question Answering (QA) tasks, and submitting runs both for the monolingual Italian and for the cross-language Italian/English tasks, the ITC-irst participation in the CLEF campaign concentrated on the Answer Validation Exercise (AVE). The participation in the AVE task, with an answer validation module based on textual entailment recognition, is motivated by our objectives of *(i)* creating a modular framework for an entailment-based approach to QA, and *(ii)* developing, in compliance with this framework, a stand-alone component for answer validation which implements different approaches to the problem.

1 Introduction

This year the ITC-irst activity on Question Answering (QA) has been concentrated on two main directions: system re-engineering and entailment-based QA. The objective of the re-engineering activity is to transform our DIOGENE QA system into a more modular architecture whose basic components are completely separated, easily testable with rapid turnaround of controlled experiments, and easily replaceable with new implementations. In this direction, our long-term ambition is to make all these components (or even the entire system) freely available in the future to the QA community for research purposes.

Adhering to this perspective, the objective of the research on entailment-based QA aims at creating a stand-alone package for Answer Validation (AV), which implements different strategies to select the best answer among a set of possible candidates. Besides our well tested statistical approach to the problem, described in [11], we plan to include in this package a new AV method based on textual entailment recognition, allowing the user for experimentations with different AV settings.

Up to date we are halfway in this ambitious roadmap, and unfortunately the results of our participation in CLEF reflect this situation. While most of the original system's components are now separated from each other, little has been done to improve their performance, and a number of bugs still affect the overall system's behavior. The main improvements concern the substitution of the MG search engine with Lucene (an open-source, cross-platform search engine,

which allows for customizable document ranking schemes), and the extension of the Answer Extraction component with a new WordNet-based module for handling “*generic*” factoid questions (*i.e.* questions whose expected answer type does not belong to the six broad named entity categories currently handled by DIOGENE). Both of these improvements, and the results achieved by DIOGENE in the Monolingual Italian and cross-language Italian/English tasks are presented in Section 5.

As for the AV problem, even though the new component based on textual entailment recognition has not been integrated in the new DIOGENE architecture, it has been separately evaluated within the CLEF-2006 Answer Validation Exercise (AVE). The description of such component, its role in the upcoming new DIOGENE architecture, together with the results achieved in the AVE task represent the focus of this paper (Sections 2, and 3, and 4).

ITC-first also provided support to the University “A.I.Cuza” of Iasi, Romania, involved in the cross-language Romanian/English QA task. Such support (concerning the document retrieval, answer extraction, and answer validation phases) is not documented in the paper as it is based on the same components used for our QA submissions.

2 Answer Validation in the QA Loop

Answer validation is a crucial step in the Question Answering loop. After a *question processing* phase, where a natural language question is analyzed in order to extract all the information (*e.g.* relevant keywords, the expected answer type) necessary for the following steps of the process, a *passage retrieval* phase (where a ranked list of relevant passages is extracted from a target collection), and an *answer extraction* phase (where possible answers are extracted from the retrieved passages), QA systems are usually required to rank a large number of answer candidates, and finally select the best one.

The problem of answer ranking/validation has been addressed in different ways, ranging from the use of *answer patterns*, to the adoption of *statistical techniques* and other *type checking* approaches based on semantics.

Pattern-based techniques (see for example [7], and [16]) are based on the use of surface patterns (regular expressions, either manually created or automatically extracted from the Web) in order to model the likelihood of answers given the question.

Statistical techniques [11] are based on computing the probability of co-occurrence of the question terms and a candidate answer either on the Web, or in a local document collection.

Semantic type checking [15] techniques exploit structured and semi-structured data sources to determine the semantic type of suggested answers, in order to retain only those matching the expected answer type category.

Most of the proposed approaches, however, are not completely suitable to model all the linguistic phenomena that may be involved in the QA process. This particularly holds when the challenge moves from factoid questions (questions for

which a complete answer can be given in a few words, usually corresponding to a named entity) to more complex ones (*e.g.* temporally restricted questions, “why questions”, “how questions”). In order to capture the deeper semantic phenomena that are involved in determining answerhood, more complex techniques that approximate the forms of inference required to identify valid textual answers become necessary. In this direction, a more complex approach has been recently proposed by [4], which experiments with different techniques based on the recognition of *textual entailment* relations, either between questions and answers, or between questions and retrieved passages, to improve the accuracy of an open-domain QA system. The proposed experiments demonstrate that considerable gains in performance can be obtained by incorporating a textual entailment recognition module both during answer processing (+11%, in terms of Mean Reciprocal Rank, over the baseline system which does not include any entailment recognition processor) and passage retrieval (+25% MRR).

Focusing on the answer validation problem, our work builds on the same hypothesis that recognizing a textual entailment relation between a question and its answer can enhance the whole QA process. The following Sections provide an overview of the problem, together with a description of the component designed for the participation to the AVE task at CLEF 2006.

3 Entailment Recognition and Answer Validation

While the language variability problem is well known in Computational Linguistics, a general unifying framework has been proposed only recently in [2]. In this framework, language variability is addressed by defining the notion of *textual entailment* as a relation that holds between two language expressions (*i.e.* a *text* T and an *hypothesis* H) if the meaning of H , as interpreted in the context of T , can be inferred from the meaning of T . The entailment relation is directional, as the meaning of one expression can entail the meaning of the other, while the opposite may not.

Recently, the task of automatically Recognizing Textual Entailment (RTE) attracted a considerable attention and has been included among the tasks of the PASCAL Challenge [3]. Given two text fragments (T and H), systems participating to the PASCAL-RTE Challenge are required to automatically determine whether an entailment relation holds between them or not. The view underlying the RTE challenge [3] is that different natural language processing applications, including Question Answering (QA), Information Extraction (IE), (multi-document) summarization, and Machine Translation (MT), have to address the language variability problem and would benefit from textual entailment in order to recognize that a particular target meaning can be inferred from different text variants.

As the following examples illustrate, the task potentially covers almost all the phenomena involved in language variability. Entailment can in fact be due to lexical variations (example 1), syntactic variations (example 2), semantic inferences (example 3), or complex combinations of all these levels.

1. T – Euro-Scandinavian media cheer Denmark v Sweden draw.
H– Denmark and Sweden tie.
2. T – Jennifer Hawkins is the 21-year-old beauty queen from Australia.
H– Jennifer Hawkins is Australia’s 21-year-old beauty queen.
3. T – The nomadic Raiders moved to LA in 1982 and won their third Super Bowl a year later.
H– The nomadic Raiders won the Super Bowl in 1982.

Due to the variety of phenomena involved in language variability, one of the crucial aspects for any system addressing RTE is the amount of knowledge required for filling the gap between T and H. The basic inference technique shared by almost all the systems participating in PASCAL-RTE is the estimation of the degree of overlap between T and H. Such overlap is computed using a number of different approaches, ranging from statistic measures like *idf*, to deep syntactic processing and semantic reasoning. Some of them are relevant or similar to our methodology, as it will be described in the rest of the paper. For instance, the system described in [6] relies on dependency parsing and extracts lexical rules from WordNet. A decision tree based algorithm is used to separate the positive from the negative examples.

Similarly, in [1] the authors describe two systems for recognizing textual entailment. The first one is based on deep syntactic processing. Both T and H are parsed and converted into a logical form. An event-oriented statistical inference engine is used to separate the TRUE from FALSE pairs. The second system is based on statistical machine translation models.

Finally, a method for recognizing textual entailment based on graph matching is described in [13]. To handle language variability, the system uses a maximum entropy co-reference classifier and calculates term similarities using WordNet.

3.1 A Distance-Based Approach for RTE

Our approach to RTE (fully described in [8], and [9]) is based on the intuition that, in order to discover an entailment relation between T and H, we must show that the whole content of H can be mapped into the content of T. The more straightforward the mapping can be established, the more probable is the entailment relation. Since a mapping can be described as the sequence of editing operations needed to transform T into H, where each edit operation has a cost associated with it, we assign an entailment relation if the overall cost of the transformation is below a certain threshold, empirically estimated on the training data. For this cost estimation purpose, we adopted the tree edit distance algorithm described in [19], and applied it to the syntactic representations (*i.e.* dependency trees) of both T and H.

Edit operations are defined at the level of single nodes of the dependency tree (*i.e.* transformations on subtrees are not allowed in the current implementation). Since the tree edit distance algorithm adopted does not consider labels on edges, while dependency trees provide them, each dependency relation R from a node A to a node B has been re-written as a complex label B-R concatenating the name

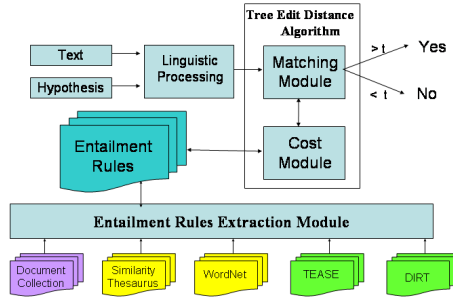


Fig. 1. RTE System architecture

of the destination node and the name of the relation. All nodes, except the root of the tree, are re-labeled in such way. The algorithm is directional: we aim to find the better (*i.e.* less costly) sequence of edit operation that transform T (also called the *source*) into H (the *target*). According to the constraints described above, the following transformations are allowed:

- **Insertion:** insert a node from the dependency tree of H into the dependency tree of T . When a node is inserted it is attached with the dependency relation of the source label.
- **Deletion:** delete a node N from the dependency tree of T . When N is deleted all its children are attached to the parent of N . It is not required to explicitly delete the children of N as they are going to be either deleted or substituted on a following step.
- **Substitution:** change the label of a node N_1 in the source tree into a label of a node N_2 of the target tree. Substitution is allowed only if the two nodes share the same part-of-speech. In case of substitution the relation attached to the substituted node is changed with the relation of the new node.

RTE System Architecture. Our RTE system is composed by the following modules, showed in Figure 1: (i) a *text processing* module, for the preprocessing of the input T/H pair; (ii) a *matching module*, which performs the mapping between T and H ; (iii) a *cost module*, which computes the cost of the edit operations.

The **text processing module** creates a syntactic representation of a T/H pair and relies on a sentence splitter and a syntactic parser. For sentence splitting we used *MXTerm* [14], a Maximum entropy sentence splitter. For parsing we used *Minipar*, a principle-based English parser [10] which has high processing speed and good precision.

The **matching module**, which implements the edit distance algorithm described in Section 3.1, finds the best sequence of edit operations between the dependency trees obtained from T and H . We define an entailment score function in the following way:

$$\gamma_{nomap}(T, H) = \gamma(\Lambda, H) + \gamma(T, \Lambda) \quad (1)$$

$$score_{entailment}(T, H) = \frac{\gamma(T, H)}{\gamma_{nomap}(T, H)} \quad (2)$$

where $\gamma(T, H)$ is the function that calculates the edit distance between T and H and $\gamma_{nomap}(T, H)$ is the *no mapping* distance equivalent to the cost of inserting the entire text of H ($\gamma(\Lambda, H)$) and deleting the entire text of T .

A similar approach is presented in [12], where the entailment score of two documents d and d' is calculated by comparing the sum of the weights (idf) of the terms that appear in both the documents to the sum of the weights of all terms in d' . We used a threshold t such that if $score(T, H) < t$ then T entails H , otherwise no entailment relation holds for the pair. To set the threshold we have used both the positive and negative examples of the training set provided by the PASCAL-RTE dataset.

The matching module makes requests to the **cost module** in order to receive the cost of each single edit operation needed to transform T into H . For this purpose, we use different cost estimation strategies, based on the knowledge acquired from different linguistic resources. In particular:

- **Insertion.** The intuition underlying insertion is that its cost is proportional to the relevance of the word w to be inserted. We measure the relevance in terms of information value (*inverse document frequency (idf)*), position in the dependency tree, and ambiguity (in terms of number of WordNet senses).
- **Deletion.** Deleted words influence the meaning of already matched words. This requires that the evaluation of the cost of a deleted word is done after the matching is finished, with measures for estimating co-occurrence (*e.g.* Mutual Information).
- **Substitution.** The cost of substituting a word w_1 with a word w_2 can be estimated considering the semantic entailment between the words. The more the two words are entailed, the less the cost of substituting one word with the other. We have defined a set of entailment rules over the WordNet relations among synsets, with their respective probabilities. If A and B are synsets in WordNet 2.0, then we derived an entailment rule in the following cases: A is hypernym of B ; A is synonym of B ; A entails B ; A pertains to B .

The difficulty of the RTE task explains the relatively poor performance of the systems participating in the PASCAL Challenge. Most of them, in fact, achieved accuracy results between 52-58% at PASCAL1 and 55-65% at PASCAL2, with a random baseline of 50%. Even though our system's performance is in line with these results, (55% at PASCAL 1 and 60.5% at PASCAL 2), our 60% accuracy on the QA pairs in the PASCAL2 test set demonstrates the potentialities of integrating RTE in a QA system, and motivated our participation in the CLEF-2006 AVE task.

3.2 RTE for Answer Validation

In terms of the approach to RTE previously described, the answer validation task can be attacked as follows:

A candidate answer ca to a question q is a *correct answer* if and only if the text from which ca is extracted entails the affirmative form of q , with ca inserted into the appropriate position.

As an example, consider the question/answer pair:

Q: “Which country did Iraq invade in 1990?”

A: “Kuwait”

In our approach, the solution to the problem consists in determining if the text T from which ca is extracted (for instance: “*The ship steamed into the Red Sea not long after Iraq invaded Kuwait in 1990.*”), entails the affirmative form H of q (i.e. “*Iraq invaded the country of Kuwait in 1990.*”).

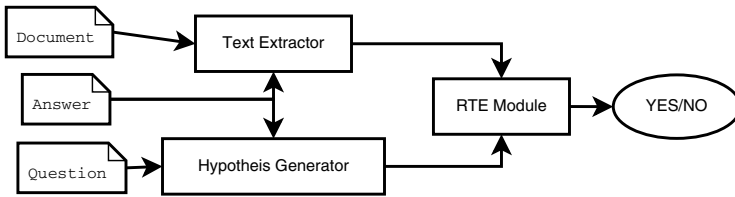


Fig. 2. AV module architecture

A possible architecture of an answer validation component based on such RTE-based approach is depicted in Figure 2. In such architecture, the edit distance algorithm is launched over the output of two pre-processing modules, namely the *Text Extractor*, and the *Hypothesis Generator*.

The former, in charge of producing T , extracts from the document containing the candidate answer several sentences that will serve as input to the RTE module. As normally only one of these sentences contains the answer, the Text Extractor should be able to deal with different discourse phenomena (such as anaphora, ellipses, etc.) before constructing the final H to be passed to the RTE module.

The latter, in charge of producing H , transforms the input question into its affirmative form and inserts the answer on the proper position. To this aim, we developed a rule-based approach which addresses the problem by converting the dependency tree of the question into a syntactic template [17].

Finally, the RTE module produces a YES/NO answer, together with a certain confidence score.

4 Participation in the AVE Task

Since the PASCAL and AVE evaluation scenarios are slightly different, the datasets provided by the organizers of the two challenges also differ. In particular,

Table 1. AVE results

Precision(YES)	Recall(YES)	F measure
0.3025	0.5023	0.3776

the input text portions (candidate answers) contained in the AVE dataset are *document fragments* extracted from the participants' submissions to the CLEF QA track. Under these conditions, the complexity of the AVE task is undoubtedly higher due to the fact that: *(i)* there are no warranties about the syntactic correctness of those fragments; *(ii)* there are no constraints about their length, and *(iii)* the answer itself is not explicitly marked within them. As a consequence, under the RTE-based framework proposed in the previous sections, the availability of a module for transforming the text portions (candidate answers) into valid T inputs for the RTE module becomes crucial.

Unfortunately we were not able to complete this module (the *Text Extractor* in Figure 2) in time for the AVE submission deadline. In order to submit a run, we used the text portions provided by the organizers as Ts without any processing. As a result, we were not able to process the 129 pairs that actually didn't contain text (but only document identifiers), and we assigned them a NO response in the final submission.

The results achieved by our system are reported in Table 1. These results are not satisfactory, as they they approximate the behavior of a random approach. As stated before, a huge drawback for our system is the missing module for converting the document fragments into valid Ts for the RTE component. The absence of such module led to a significant drop in the performance of the parser, which is a crucial component in our RTE-based architecture. With these circumstances we can not draw definitive conclusions about the performance of our system, which requires at least a grammatically correct input. We assume that introducing the missing module we will significantly improve the performance of the system, allowing to reach results comparable with the ones obtained in the PASCAL 1 & 2 challenges.

5 Participation in the QA Task: Towards a New DIOGENE System

Our participation in the CLEF-2006 QA task is the first step on the way of developing a new and improved DIOGENE system. The leading principle of this re-engineering activity is to create a modular architecture, open to the insertion/substitution of new components. Another long-term objective of our work on QA, is to make the core components of the system freely available to the QA community for research purposes. In this direction, new implementation guidelines have been adopted, starting from the substitution of Lisp with Java as the main programming language. As stated before, our choices we were led by the following objectives:

1. Modularity - Re-build the old monolithic system into a pipeline of modules which share common I/O formats (xml files).
2. Configurable - Allow for the capability of configuring the settings of the different modules with external configuration files. We have developed a common xml schema for configuration description.
3. Evaluation - Provide the capability of performing fine-grained evaluation cycles over the individual processing modules which compose a QA system. In order to make a step-to-step evaluation of our modules we started to build reference datasets. Having such, the results of each processing step can be compared against them. At this stage, we focus on factoid questions from the TREC-2002 and TREC-2003 testsets. Exploiting such data, for instance, we are now able to perform separate tests on the Document Retrieval and the Candidate Answer Extraction steps. In the new architecture, evaluation modules are completely detached from processing modules. The former are independent from specific datasets and/or formats, enabling a quick porting along with the evolution of the task.

In spite of the huge effort in the modularity direction (now all the system's modules have been completely detached), little has been done to improve the behavior of the core components. The only improvements over the system described in [17] concern the *document retrieval* and *answer extraction* phases.

As for **document retrieval**, this year's novelty is the substitution of the MG search engine [18] with Lucene [5]. The many advantages of Lucene now do enable us to implement more refined strategies for indexing, ranking, and querying the underlying text collection. The Document Retrieval Evaluation is currently performed against the previous TREC/QA test sets, for which correct results are available. This only allows for a rough evaluation, since NIST results only report actual answers given by competing systems. However, for our goals such collections provide a reasonably fair coverage of correct answers. We perform Document Retrieval Evaluation at two different levels: Document Level and Text Passage Level. The first aims at evaluating the system against the official task, in which only the document reference must be provided. Conversely, Text Passage Level evaluation provides us with a better perspective on the following processing step, which is Candidate Answer Extraction. In fact, we fragment the AQUAINT corpus in paragraphs (our Text Passages) with respect to indexing. And, these paragraphs are exactly the text units over which answer extraction is performed.

As for **answer extraction**, we improved the candidate answers' selection process by developing a more effective way to handle "*generic*" factoid questions (*i.e.* questions whose expected answer type does not belong to the six broad named entity categories previously handled by DIOGENE). Generic questions are now handled extracting as possible answer candidates all the hyponyms of the question focus that are found in the text passages returned by the search engine. For instance, given the question Q: "*What instrument did Jimi Hendrix play?*", the answer extraction module will select "*guitar*" as a possible answer candidate as it is an hyponym of "*instrument*", the focus of the question.

Table 2. System performance in the QA tasks

<i>task</i>	<i>Overall (%)</i>	<i>Def. (%)</i>	<i>Factoid (%)</i>	<i>Temp. (%)</i>
Italian/Italian	22.87	17.07	25.00	0.00
Italian/English	12.63	0.00	16.00	0.00

Apart from these modifications, the runs submitted to this year's edition of the CLEF QA task (results are reported in Table 2) have been produced with the old system's components, and reflect the "work-in-progress" situation of DIOGENE.

References

1. Bayer, S., Burger, J., Ferro, L., Henderson, J., Yeh, A.: Mitre's submissions to the eu pascal rte challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 309–331. Springer, Heidelberg (2006)
2. Dagan, I., Glickman, O.: Generic applied modeling of language variability. In: Proceedings of PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble (2004)
3. Dagan, I., Glickman, O., Magnini, B.: The pascal recognizing textual entailment challenge. In: Proceedings of PASCAL Workshop on Recognizing Textual Entailment, Southampton, UK (2005)
4. Harabagiu, S., Hickl, A.: Methods of using textual entailment in open-domain question answering. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics, ACL-2006, Sydney Australia (July 2006)
5. Hatcher, E., Gospodnetic, O.: Lucene in Action (In Action series). Manning Publications (December 2004)
6. Herrera, J., Penas, A., Verdejo, F.: Textual entailment recognition based on dependency analysis and wordnet. In: Proceedings of PASCAL Workshop on Recognizing Textual Entailment, Southampton, UK (2005)
7. Hovy, E., Hermjakob, U., Ravichandran, A.D.: Question/answer typology with surface text patterns. In: Proceedings of the DARPA Human Language Technology Conference (HLT), San Diego, CA (2002)
8. Kouylekov, M., Magnini, B.: In: Proceedings of PASCAL Workshop on Recognizing Textual Entailment, Southampton, UK (2005)
9. Kouylekov, M., Magnini, B.: In: Proceedings of PASCAL Workshop on Recognizing Textual Entailment, Venezia, UK (2006)
10. Lin, D.: A maximum entropy part-of-speech tagger. In: Proceedings of the Workshop on Evaluation of Parsing Systems at LREC-98, Granada, Spain (1998)
11. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Is it the right answer? exploiting web redundancy for answer validation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002, Philadelphia (PA), 7–12 July 2002, pp. 1495–1500 (2002)
12. Monz, C., de Rijke, M.: Light-weight entailment checking for computational semantics. In: Proceedings of The third workshop on inference in computational semantics, Siena, Italy, pp. 59–72 (2001)

13. Raina, R., Haghighi, A., Cox, C., Finkel, J., Michels, J., Toutanova, K., MacCartney, B., de Marneffe, M.-C., Manning, C.D., Ng, A.Y.: Robust textual inference using diverse knowledge sources. In: *Proceedings of PASCAL Workshop on Recognizing Textual Entailment*, Southampton, UK (2005)
14. Ratnaparkhi, A.: A maximum entropy part-of-speech tagger. In: *Proceeding of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania (1996)
15. Schlobach, S., Olsthoorn, M., de Rijke, M.: Type checking in open-domain question answering. In: *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)* (2004)
16. Subbotin, M.: Patterns of potential answer expressions as clues to the right answers. In: *Proceedings of the TREC-10 Conference*, pp. 175–182 (2001)
17. Tanev, H., Kouylekov, M., Negri, M., Magnini, B.: Exploiting linguistic indices and syntactic structures for multilingual question answering: Itc-irst at clef 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
18. Witten, I.H., Moffat, A., Bell, T.C.: *The second edition of Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, San Francisco (1999)
19. Zhang, K., Shasha, D.: Fast algorithm for the unit cost editing distance between trees. *Journal of algorithms* 11, 1245–1262 (1990)

Link-Based vs. Content-Based Retrieval for Question Answering Using Wikipedia

Sisay Fissaha Adafre, Valentin Jijkoun, and Maarten de Rijke

ISLA, University of Amsterdam
{sfissaha,jijkoun,mdr}@science.uva.nl

Abstract. We describe our participation in the WiQA 2006 pilot on question answering using Wikipedia, with a focus on comparing link-based vs content-based retrieval. Our system currently works for Dutch and English.

1 Introduction

The WiQA (Question Answering using Wikipedia) task derives much of its motivation from the observation that, in the Wikipedia context, the distinction between reader and author has become blurred; see [3] for an overview of the task. Given a topic (and associated article) in one language, identify relevant snippets on the topic from other articles in the same language or even in other languages. In the context of Wikipedia, having a system that offers this type of functionality capability is important for readers as well as authors.

In this report, we describe our participation in the WiQA 2006 pilot. We took part both in the monolingual and bilingual tasks. Our approach consists of three steps: retrieving candidate snippets, reranking the snippets, and removing duplicates. We compare two approaches for identifying candidate snippets, i.e., link-based retrieval and traditional content-based retrieval. Section 2 presents an overview of the system; Section 3 describes our runs; Section 4 provides our results, and Section 5 presents conclusions.

2 System Description

We briefly describe our mono- and bilingual systems for the WiQA 2006 pilot. The main components of the monolingual system are aimed at addressing the following subtasks: identifying relevant snippets, estimating sentence importance, and removing redundant snippets. We briefly discuss each of these in turn.

We apply two approaches to identifying relevant sentences. The first exploits the peculiar characteristics of Wikipedia, i.e., the dense link structure, to identify relevant snippets. The link structure, and in particular, the structure of the incoming links (similar to Wikipedia’s “What links here” feature), provides a simple mechanism for identifying relevant articles. If an article contains a citation to the topic then it is likely to contain relevant bits of information about the topic

since the hyperlinks in Wikipedia are part of the content of the article. Since hyperlinks are created manually by humans, this approach tends to produce little noise. However, not all mentions of a topic may be hyperlinked which may cause some recall problems. Furthermore, coreferences may also need to be resolved in order to improve recall. Our second approach to identifying relevant sentences is based on an out-of-the-box Lucene [4] index of Wikipedia articles. From retrieved articles, we extract sentences that mention (sufficiently large parts of) the topic for which we are attempting to locate relevant information. We expect the latter content-based approach to be more recall-oriented than the link-based approach, potentially picking up more noise in the process.

For estimating sentence importance, we combine several types of evidence: retrieval score, position in the article, and a graph-based ranking score. We normalize the retrieval scores (obtained in the first step) to values between 0 and 1 by dividing each retrieval score with the maximum value. Furthermore, the earlier a sentence appears in an article, the more important we assume it to be. Sentence positions are converted into values between 0 and 1 in which the sentence at position 1 receives the maximum graph-based score (explained below), and subsequent sentences receive a score which is a fraction of the maximum graph-based score using the formula proposed in [5]. Finally, graph-based scoring allows us to rank sentences by taking into account the more global context of sentences, which consists of a representative sample of articles that belong to the same categories as the topic. The resulting corpus serves as our importance model by which we assign each sentence a score; see [2].

For the final filtering step, we compare each candidate sentence with the sentences in the main article, and sentences that appear before it in the list of sentences ranked by importance score. We used simple word overlap for measuring sentence similarity. We keep the maximum similarity score and subtract it from the sentence score. We sort the list again in decreasing order of the resulting scores, and return the top 10 sentences as output.

As to our approach to the multilingual (Dutch-English) task, given a topic in one language, we need to find important snippets in other Wikipedia articles of the same language or other languages. The main challenge is to ensure relevance and novelty across languages. To measure similarity among snippets from different languages we used a bilingual dictionary generated from Wikipedia itself (from the titles of corresponding articles in the two languages) and applied the method to Dutch and English articles. Briefly, given a topic in one language, we identify important snippets in the language of the topic translate the topic into other languages (using the bilingual dictionary), identify important snippets in the translated topics, and filter the resulting multilingual list for redundant information, again using the bilingual dictionary; see [1].

3 Runs

We submitted a total of seven runs: six monolingual runs for Dutch and English (three runs per language), and one is a Dutch-English bilingual run. All the

monolingual runs use the approach outlined in Section 2. Except for the use of stopword lists, the method is completely generic and can be ported to any language with relative ease. The runs differ in the methods adopted for acquiring the initial set of relevant sentences: link-based (*Link*), content-based (*Ret*), or a combination of the two (*LinkRet*).

For the bilingual run we consider Dutch and English articles. The source topics can be in any of these languages. As mentioned in the previous section, we build on top of our monolingual approach. We used the output of the third monolingual run as an input for the bilingual filtering.

4 Results

We present the results of our seven runs. The results are assessed based on the following attributes: “supported,” “importance,” “novelty,” and “non-repetition.” The summary statistics that we report are *Avg. Yield* (the total number of supported, important, novel, non-repeated snippets in the top 10, per topic, averaged over all topics); *MRR* (mean reciprocal rank of the first supported, important, novel, non-repeated snippet in the top 10; *p@10* (the fraction of supported, important, novel, non-repeated snippets in the top 10 snippets, per topic, averaged over all topics). See the overview paper for the WiQA pilot task for further details [3]. Table 1 shows the results of the seven runs.

The scores for the *Link* based monolingual runs are the best, suggesting that link-based retrieval provides a more accurate initial set of relevant sentences on which the performance of the whole system largely depends. The content-based approach seems to introduce more noise as shown by the scores of the *Ret* and *LinkRet* based monolingual runs for both languages. Contrary to our expectation, the combination of the two methods performed worse for English.

In order to see whether our approaches return similar sets of snippets, we compared the result sets, i.e. sets of supported, important, novel and non-repeated snippets returned by the methods. While *Link* returned 353 *important* snippets and *Ret* returned 325 snippets in top 20, the joint pool of the two methods contains 484 important snippets. This indicates that there is a potential room for improvement in combining the link-based and retrieval-based methods for importance estimation. Unfortunately, our combination method used in the run *LinkRet* did not perform well. In future work we will focus on other ways to effectively combine different sources of evidence for this task.

Table 1. Results for the English and Dutch monolingual tasks and the Dutch-English bilingual task

English				Dutch			English-Dutch		
	Avg.	Yield	MRR p@10	Avg.	Yield	MRR p@10	Avg.	Yield	MRR p@10
Ret	2.938	0.523	0.329	3.200	0.459	0.427			
Link	3.385	0.579	0.358	3.800	0.532	0.501			
LinkRet	2.892	0.516	0.330	3.500	0.532	0.494	5.03	0.518	0.535

The scores for the bilingual run tend to be higher than our monolingual scores. This may be due to the fact that most of the topics are Dutch topics that are mostly short and additional snippets are likely to be new. Furthermore, the snippets can come from both Dutch and English encyclopedias which also contributes to finding good snippets.

5 Conclusion

We have described our participation in the WiQA 2006 pilot task. Our approach consists of three steps: retrieving candidate snippets, reranking the snippets, and removing duplicates. We compared two approaches for identifying candidate snippets: link-based retrieval and traditional content-based retrieval. The results show that the link-based approach performed better than the content-based approach, and although a combined approach showed a relatively poor performance, our analysis suggests potential room for improvement in combining the two methods.

Acknowledgments

Sisay Fissaha Adafre was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001. Valentin Jijkoun was supported by NWO under project numbers 220-80-001, 600.-065.-120 and 612.000.106. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.-000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

- [1] Fissaha Adafre, S., de Rijke, M.: Finding similar sentences across multiple languages in wikipedia. In: EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources (2006)
- [2] Fissaha Adafre, S., de Rijke, M.: Learning to identify important biographical facts (Submitted)
- [3] Jijkoun, V., de Rijke, M.: Overview of WiQA 2006 (In This volume) (2006)
- [4] Lucene: The Lucene search engine, <http://lucene.apache.org/>
- [5] Radev, D.R., Jing, H., Sty, M., Tam, D.: Centroid-based summarization of multiple documents. *Information Processing and Management* 40(6), 919–938 (2004)

Identifying Novel Information Using Latent Semantic Analysis in the WiQA Task at CLEF 2006

Richard F.E. Sutcliffe¹, Josef Steinberger², Udo Kruschwitz³,
Mijail Alexandrov-Kabadjov³, and Massimo Poesio³

¹ Documents and Linguistic Technology Group, Department of Computer Science
and Information Systems, University of Limerick, Limerick, Ireland

Richard.Sutcliffe@ul.ie

² Department of Computer Science and Engineering, University of West Bohemia,
Univerzitni 8, 306 14 Plzen, Czech Republic

jstein@kiv.zcu.cz

³ Department of Computer Science, University of Essex,
Wivenhoe Park, Colchester CO4 3SQ, UK

{udo, malexa, poesio}@essex.ac.uk

Abstract. In our two-stage system for the English monolingual WiQA Task, snippets were first retrieved if they contained an exact match with the title. Candidates were then passed to the Latent Semantic Analysis component which judged them Novel if their match with the article text was less than a threshold. In Run1, the ten best snippets were returned and in Run 2 the twenty best. Run 1 was superior, with Average Yield per Topic 2.46 and Precision 0.37. Compared to other groups, our performance was in the middle of the range except for Precision where our system was the best. We attribute this to our use of exact title matches in the IR stage. In future work we will vary the approach used depending on the topic type, exploit co-references in conjunction with exact matches and make use of the elaborate hyperlink structure which is a unique and most interesting aspect of the Wikipedia.

1 Introduction

This article outlines an experiment in the use of Latent Semantic Analysis (LSA) for selecting information relevant to a topic. It was carried out within the Question Answering using Wikipedia (WiQA) Pilot Task which formed part of the Multiple Language Question Answering Track at the 2006 Cross Language Evaluation Forum (CLEF).

The Wikipedia [1] is a multilingual free-content encyclopaedia which is publicly accessible over the Internet. From the perspective of WiQA it can be viewed as a set of articles each with a unique title. During their preliminary work, the task organisers created an XML compliant corpus from the English Wikipedia articles [2]. The title of each article was assigned automatically to one of four subject categories PERSON, LOCATION, ORGANIZATION and NONE. At the same time the text of each article was split into separate sentences each with its own identifier. The complex hyperlink structure of the original Wikipedia is faithfully preserved in the corpus, although we did not use this in the present project.

Table 1. Training Data. A sample topic together with the corresponding article text and three candidate snippets. Topic titles are unique in Wikipedia. In our system, the title was added to the start of the article which is why the name appears twice. The existence of the double quotes is connected with the removal of hyperlinks – an imperfect process. Are the example snippets Important and Novel or not? See the text for a discussion.

Topic	Carolyn Keene
Article	Carolyn Keene Carolyn Keene is the pseudonym of the authors of the Nancy Drew mystery series, published by the Stratemeyer Syndicate. Stratemeyer hired writers, including Mildred Benson, to write the novels in this series, who initially were paid only \$125 for each book and were required by their contract to give up all rights to the work and to maintain confidentiality. Edward Stratemeyer's daughter, Harriet Adams, also wrote books in the Nancy Drew series under the pseudonym. Other ghostwriters who used this name to write Nancy Drew mysteries included Leslie McFarlane, James Duncan Lawrence, Nancy Axelrod, Priscilla Doll, Charles Strong, Alma Sasse, Wilhelmina Rankin, George Waller Jr., Margaret Scherf, and Susan Wittig Albert. " " by John Keeline lists the ghostwriters responsible for some individual Nancy Drew books.
Snippet 1	The name Carolyn Keene has also been used to author a shorter series of books entitled The Dana Girls.
Snippet 2	All Nancy Drew books are published under the pseudonym Carolyn Keene regardless of actual author.
Snippet 3	Harriet Adams (born Harriet Stratemeyer , pseudonyms Carolyn Keene and Franklin W. Dixon) (1893 - 1982), U.S. juvenile mystery novelist and publisher; wrote Nancy Drew and Hardy Boys books.

The general aim of WiQA this year was to investigate methods of identifying information on a topic which is present somewhere in the Wikipedia but not included in the article specifically devoted to that topic. For example, there is an article entitled ‘Johann Sebastian Bach’ in the Wikipedia. The question WiQA sought to answer is this: What information on Bach is there within the Wikipedia *other than* in this article? The task was formalised by providing participants with a list of articles and requiring their systems to return for each a list of up to twenty sentences (henceforth called snippets) from other articles which they considered relevant to the article and yet not already included in it. There were 65 titles in the test set, divided among the categories PERSON, LOCATION, ORGANIZATION and NONE. Evaluation of each snippet was on the basis of whether it was **supported** (in the corpus), **important** to the topic, **novel** (not in the original article) and **non-repeated** (not mentioned in

previously returned snippets for this topic). Evaluation of systems was mainly in terms of the snippets judged supported and important and novel and non-repeated within the first ten snippets returned by a system for each topic. A detailed description of the task and its associated evaluation measures can be found in the general article on the WiQA task in this volume.

Latent Semantic Indexing (LSI) was originally developed as an Information Retrieval (IR) technique [3]. A term-by-document matrix of dimensions $t \times d$ of the kind commonly used for inverted indexing in Information Retrieval (IR) is transformed by Singular Value Decomposition into a product of three matrices $t \times r$, $r \times r$ and $r \times d$. The $r \times r$ matrix is diagonal and contains the eponymous ‘singular values’ in such a way that the top left element is the most important and the bottom right element is the least important. Using the original $r \times r$ matrix and multiplying the three matrices together results exactly in the original $t \times d$ matrix. However, by using only the first n dimensions ($1 \leq n \leq r$) in the $r \times r$ matrix and setting the others to zero, it is possible to produce an approximation of the original which nevertheless captures the most important common aspects by giving them a common representation. In the original IR context, this meant that even if a word was not in a particular document it could be detected whether or not another word with similar meaning was present. Thus, LSI could be used to create representations of word senses automatically.

In abstract terms, LSI can be viewed as a method of squeezing information through a bottleneck in such a way that common aspects of the data must be discovered along the way. Following on from the original work it was realised that this idea was applicable to a wide range of tasks including information filtering [4] and cross-language information retrieval [5]. Outside IR, the technique is usually referred to as Latent Semantic Analysis (LSA). Within NLP, LSA has been applied to a variety of problems such as spelling correction [6], morphology induction [7], text segmentation [8], hyponymy extraction [9], summarisation [10], and noun compound disambiguation, prepositional phrase attachment and coordination ambiguity resolution [11].

In the following sections we first describe how we created test data for the project and used it to develop our algorithm. We then summarise the final architecture of the system before stating the runs which were carried out and analysing the results obtained. Finally, we draw some conclusions and make recommendations for further work.

2 Algorithm Development

We decided on a very simple form of algorithm for our experiments. In the first stage, possibly relevant snippets (i.e. sentences) would be retrieved from the corpus using an IR system. In the second, these would be subjected to the LSA technique in order to estimate their novelty. In previous work, LSA had been applied to summarisation by using it to decide which snippets were related to a topic [10]. In this project, the idea was reversed by trying to establish that snippets were novel on the basis that they were **not** ‘related’ to the original article.

As this was the first year of the task, there was no training data. However, the organisers did supply 80 example topics. Twenty of these were selected. For each, the

title was submitted as an exact IR search to a system indexed by sentence on the entire Wikipedia corpus. The ‘documents’ (i.e. snippets) returned were saved, as was the bare text of the original article. The snippets were then studied by hand and by reference to the original document were judged to be either ‘relevant and novel’ or not. This data was then used in subsequent tuning.

An example training topic (Carolyn Keene) can be seen in Table 1. Below it are three sample snippets returned by the IR component because they contain the string ‘Carolyn Keene’. The first one is clearly Important and Novel because it gives information about a whole series of books written under the same name and not mentioned in the article. The second one is Important because it states that all Nancy Drew books are written under this name. However, it is not Novel because this information is in the original article. Now, consider the third example concerning Harriet Adams. The decision here is not quite so easy to make. Adams is mentioned in the article, so the fact that she wrote some of the books is not Novel. However, there is other information which is Novel, for example that she wrote books in the Hardy Boys series and also that she had another pseudonym, Franklin W. Dixon. The question is whether such information is Important. This is very hard to judge. Therefore we concluded that the task was not straightforward even for humans.

3 System Architecture and Implementation

Following our analysis of the training data, it was decided to adopt a similar approach in the final system. This involved retrieving snippets by exact phrase match. To facilitate this process, the corpus was re-formatted replacing sentence and document identifiers within attributes by the equivalent in elements and at the same time removing all hyperlink information which can occur within words and thus affects retrieval. The new version of the corpus was then indexed using Terrier [12,13] with each individual snippet (sentence) being considered as a separate document. This meant that any matching of an input query was entirely within a snippet and never across snippets.

An input query consists of the title of a Wikipedia article. An exact search is performed using Terrier, resulting in an ordered list of matching snippets. Those coming from the original article are eliminated. In the actual runs, the number of snippets varies from 0 (where no matching snippets were found at all) to 947. The data is then formatted for LSA processing. The bare text of the original article with no formatting and one sentence per line, including the title which forms the first line, is placed in a text file. A second file is prepared for the topic which contains the snippets, one snippet per line. This file therefore contains between 0 and 947 lines, depending on the topic. LSA processing is then carried out. This assigns to each snippet a probability as to whether it is novel with respect to the topic or not. The n snippets with highest probability are then returned, preserving the underlying order determined by Terrier. n is either 10 or 20 depending on the run. In the final stage, an XML document is created listing for each topic the selected snippets.

Table 2. Results. Both runs were identical except that in Run 1 the maximum number of snippets returned by the system was 10 while in Run 2 it was 20. The values under the column **All** are those returned by the organisers. All other columns are analyses on different subsets of the topics. The object is to see if performance of the system varies by topic type. The column **Person** shows results just for topics which were of type PERSON and similarly for the columns **Location**, **Organization** and **None**. **Empty** denotes just those topics which contain no text at all (!) while **Long** is restricted to ‘long’ topics which contain a lot of text.

Run 1							
	All	Pers	Loc	Org	None	Empty	Long
No. Topics	65	16	18	16	15	6	15
No. Snippets	435	113	119	112	91	33	123
Supported Snippets	435	113	119	112	91	33	123
Important	226	74	54	61	37	13	49
Important & Novel	165	54	37	51	23	13	29
Important & Novel & Non-Repeated	161	52	37	49	23	12	29
Yield (Top 10)	160	52	36	49	23	12	29
Avg. Yield per Topic (Top 10)	2.46	3.25	2.00	3.06	1.53	2.00	1.93
Mean Reciprocal Rank	0.54	0.67	0.47	0.66	0.34	0.56	0.59
Precision (Top 10)	0.37	0.46	0.30	0.44	0.25	0.36	0.24
Run 2							
	All	Pers	Loc	Org	None	Empty	Long
No. Topics	65	16	18	16	15	6	15
No. Snippets	682	155	206	188	133	59	210
Supported Snippets	682	155	206	188	133	59	210
Important	310	87	93	84	46	26	81
Important & Novel	223	60	60	72	31	26	45
Important & Novel & Non-Repeated	194	52	53	61	28	19	36
Yield (Top 10)	152	44	37	47	24	15	24
Avg. Yield per Topic (Top 10)	2.34	2.75	2.06	2.94	1.60	2.50	1.60
Mean Reciprocal Rank	0.50	0.58	0.47	0.59	0.36	0.50	0.56
Precision (Top 10)	0.33	0.39	0.29	0.39	0.26	0.38	0.19

4 Runs and Results

Two runs were submitted, Run 1 in which the number of accepted snippets was at maximum 10, and Run 2 in which it was at maximum 20. In all other respects the runs were identical.

Table 2 summarises the overall results. In addition to the official figures returned by the organisers (denoted All) we also computed results on various different subsets of the topics. The objective was to see whether the system performed better on some types of topic than on others. Each topic in the corpus had been automatically assigned to one of four categories by the organisers using Named Entity recognition tools. The categories are Person, Location, Organization and None. We decided to make use of these in our analysis. Person denotes topics categorised as describing a person and similarly for Location and Organization. None is assigned to all topics not considered to be one of the first three. To these four were added two further ones of our own invention. Empty denotes an interesting class of topics which consist of a title and no text at all. These are something of an anomaly in the Wikipedia, presumably indicating work in progress. In this case the WiQA task is effectively to create an article from first principles. Finally, Long denotes lengthy articles, rather generally defined by us as ‘more than one page on the screen’.

Each snippet was judged by the evaluators along a number of dimensions. A Supported snippet is one which is indeed in the corpus. If it contains significant information relevant to the topic it is judged Important. This decision is made completely independently of the topic text. It is Novel if it is Important and in addition contains information not already present in the topic. Finally, it is judged Non-Repeated if it is Important and Novel and has not been included in an earlier Important and Novel snippet. Repetition is thus always judged between one snippet and the rest, not relative to the original topic text.

The main judgements in the evaluation are relative to the top ten snippets returned for a topic. The Yield is the count of Important, Novel and Non-Repeated snippets occurring in the first ten, taken across all topics in the group under consideration (e.g. All). The Average Yield is this figure divided by the number of topics in the group, i.e. it is the Yield per topic. Reciprocal Rank is the inverse of the position of the first Important, Novel and Non-Repeated snippet in the top ten (numbered from 1 up to ten) returned in response to a particular topic. The Mean Reciprocal Rank (MRR) is the average of these values over all the topics in the group. MRR is an attempt to measure ‘how high up the list’ useful information starts to appear in the output. Finally Precision is the mean precision (number of Important, Novel and Non-Repeated snippets returned in the first ten for a topic, divided by ten) computed over all the topics.

Of the two runs we submitted, Run 1 gave better results than Run 2 in terms of overall Average Yield per Topic, MRR and Precision at top 10 snippets. Having said that, the second run did retrieve far more snippets than the first one (682 as opposed to 435); it also retrieved more Important, Important and Novel, and Important, Novel and Non-Repeated snippets than the first run. However, what counts is how many Important, Novel and Non-Repeated snippets were found in the ten snippets that were ranked highest for each topic (i.e. the Yield). When we look at all 65 topics, then the yield was lower for Run 2 (152 vs. 160). This is not just true for the overall figures,

but also for the topics in categories Person, Organization and Long. For Location, None and Empty, yield in the second run was marginally better.

The difference in performance between the two runs can be accounted for by an anomaly in the architecture of our system. The underlying order of snippets is determined by their degree of match with the IR search query. Because this query is simply the title and nothing else, and since all returned snippets must contain the exact title, the degree of match will be related only to the length of the snippet – short snippets will match more highly than long ones. When this list of snippets is passed to the LSA component, a binary decision is made for each snippet (depending on its score) as to whether it is relevant or not. This depends on a snippet's LSA score but not on its position in the ranking. The difference between the runs lies in the number of snippets LSA is permitted to select. In Run 1 this consists of the best ten. In Run 2, when we come to select the best twenty, this is a superset of the best ten, but it may well be that lower scoring snippets are now selected which are higher in the underlying Terrier ranking than the higher scoring snippets already chosen for Run 1. In other words, our strategy could result in high scoring snippets in Run 1 being inadvertently pushed down the ranking by low scoring ones in Run 2. This effect could explain why the amount of relevant information in Run 2 is higher but at the same time the Precision is lower.

To find out where our approach works best we broke down the total number of topics into groups according to the categories identified earlier, i.e. Person (16 topics), Location (18), Organization (16), and None (15). A separate analysis was also performed for the two categories we identified, i.e. Empty (6) and Long (15). Instead of analysing individual topics we will concentrate on the aggregated figures that give us average values over all topics of a category.

We achieved the highest average Yield per Topic (3.25), highest MRR (0.67) as well as highest Precision at top 10 snippets (0.46) for topics of category Person in Run 1. In other words, for queries about persons a third of the top ten retrieved snippets were considered Important, Novel and Non-Repeated. However, the runs did not necessarily retrieve ten snippets for each query; usually we retrieved fewer than that. The Precision indicates that nearly half of all retrieved snippets were high quality matches. These values are better than what we obtained for any of the other categories (including Empty and Long) over both runs. This suggests that our methods work best at identifying Important, Novel and Non-Repeated information about persons.

We also observe that in Run 1 the Average Yield per Topic, MRR and Precision for topics of type Person or Organization are all better than those for type All. The same is true for Run 2.

On the other hand, both our runs are much worse on topics of type None. All the considered measures score much lower for topics of this category: Average Yield per Topic is 1.53, MRR is 0.34, and Precision at top 10 snippets is 0.25. Interestingly, these lowest values were recorded in Run 1 which overall is better than Run 2. Nevertheless, Precision at top 10 snippets is even lower for Long topics in both runs (0.24 and 0.19, respectively). The last figure is interesting, because this lowest Precision for the long documents in Run 2 corresponds with the lowest Average Yield per Topic (1.6) but the highest average number of snippets per topic (14.0, i.e. 210 snippets divided by 15 topics). No other category retrieved that many responses per query on average. As a comparison, the lowest average number of returned snippets (5.5, i.e. 33

snippets divided by 6 topics) was recorded for Empty topics in Run 1. The conclusion is that the system (perhaps not surprisingly) seems to suggest few snippets for short documents and far more for long documents. More returned snippets do not, however, mean better quality.

Topics classified as Long or None are therefore a major weakness of our approach, and future work will need to address this. One possibility is that we could use the topic classifications at run time and then apply different methods for different categories.

5 Conclusions

This was our first attempt at this task and at the same time we used a very simple system based around LSA applied only to snippets containing exactly the topic's title. In this context the results are not bad. The system works best for topics of type Person and Organization. Our highest precision overall was 0.46 for Persons in Run 1. Compared with the overall results of all participants, we are broadly in the middle of the range. The exception is overall Precision for Run 1 where we appear to be the highest overall in the task, with a value of 0.37. This result is probably due to our use of exact matches to titles in the IR stage of the system.

Concerning the general task, it was very interesting but in some ways it raised more questions than answers. While it is quite easy to judge Novel and Non-Repeated it is not easy to judge Important. This is a subjective matter and can not be decided upon without considering such factors as maximum article length, intended readership, the existence of other articles on related topics, hyperlink structure (e.g. direct links to other articles containing 'missing' information) and editorial policy.

We prepared our own training data and due to lack of time this only amounted to twenty topics with judged snippets. In future years there will be much more data to carry out tuning and this might well affect results.

Our policy of insisting on an exact match of a snippet with the title of a topic resulted in the vast majority of cases in the snippet being about the topic. (There are relatively few cases of topic ambiguity although a few were found.) In other words, the precision of the data passed on to the LSA component was high. On the other hand, we must have missed many important snippets. For example we used no co-reference resolution which might well have increased recall while not affecting precision, certainly in cases like substring co-reference within an article (e.g. 'Carolyn Keene ... Mrs. Keene').

The link structure of the Wikipedia is very complex and has been faithfully captured in the corpus. We did not use this at all. It would be very interesting to investigate snippet association measures based on the 'reachability' of a candidate snippet from the topic article and to compare the information they yield with that provided by LSA. However, the link structure is not all gain: In many cases only a substring of a token constitutes the link. When the markup is analysed it can be very difficult to recover the token accurately – either it is wrongly split into two or a pair of tokens are incorrectly joined. This must affect the performance of the IR and other components though the difference in performance caused may be slight.

References

1. Wikipedia (2006), <http://en.wikipedia.org>
2. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. *SIGIR Forum* 40(1) (2006)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 391–407 (1990)
4. Foltz, P.W., Dumais, S.T.: Personalized information delivery: An analysis of information filtering methods. *Communications of the Association for Computing Machinery* 35, 51–60 (1992)
5. Littman, M.L., Dumais, S.T., Landauer, T.K.: Automatic cross-language information retrieval using Latent Semantic Indexing. In: Grefenstette, G. (ed.) *Cross Language Information Retrieval*, pp. 51–62. Kluwer Academic Publishers, Norwell, MA (1998)
6. Jones, M.P., Martin, J.H.: Contextual spelling correction using Latent Semantic Analysis. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP '97)*, pp. 166–173 (1997)
7. Schone, P., Jurafsky, D.: Knowledge-free induction of morphology using Latent Semantic Analysis. In: *Proceedings of the Fourth Conference on Computational Natural Language Learning (CoNLL-2000) and the Second Learning Language in Logic Workshop (LLL-2000)*, pp. 67–72 (2000)
8. Choi, F.Y.Y., Wiemer-Hastings, P., Moore, J.D.: Latent Semantic Analysis for Text Segmentation. In: *Proceedings of EMNLP, Pittsburgh* (2001)
9. Cederberg, S., Widdows, D.: Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: *Proceedings of the Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*, pp. 111–118 (2003)
10. Steinberger, J., Kabadjov, M.A., Poesio, M., Sanchez-Graillet, O.: Improving LSA-based Summarization with Anaphora Resolution. In: *Proceedings of Human Language Technology Conference / Conference on Empirical Methods in Natural Language Processing*, Vancouver, Canada, pp. 1–8 (October 2005)
11. Buckeridge, A. M.: Latent Semantic Indexing as a Measure of Conceptual Association for the Unsupervised Resolution of Attachment Ambiguities. Ph.D. Thesis, University of Limerick (2005)
12. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, Seattle, Washington, USA (August 10, 2006)
13. Terrier (2006), <http://ir.dcs.gla.ac.uk/terrier/>

A Bag-of-Words Based Ranking Method for the Wikipedia Question Answering Task

Davide Buscaldi and Paolo Rosso

Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi,proso}@dsic.upv.es

Abstract. This paper presents a simple approach to the Wikipedia Question Answering pilot task in CLEF 2006. The approach ranks the snippets, retrieved using the Lucene search engine, by means of a similarity measure based on bags of words extracted from both the snippets and the articles in wikipedia. Our participation was in the monolingual English and Spanish tasks. We obtained the best results in the Spanish one.

1 Introduction

Question Answering (QA) based on Wikipedia (*WiQA*) is a new exercise, which was proposed as a pilot task in CLEF 2006. Wikipedia recently caught the attention of various researchers [6,1] as a resource for the QA task, in particular for the direct extraction of answers. *WiQA* is a quite different task, since it is aimed at helping the readers/authors of Wikipedia rather than finding answers to user questions. In the words of the organizers¹, the purpose of the *WiQA* pilot is “to see how IR and NLP techniques can be effectively used to help readers and authors of Wikispaces get access to information spread throughout Wikipedia rather than stored locally on the pages”. An author of a given Wikispaces can be interested in collecting information about the topic of the page that is not yet included in the text, but is relevant and important for the topic, so that it can be used to update the content of the Wikispaces. Therefore, an automatic system will provide the author with information snippets extracted from Wikipedia with the following characteristics:

- *unseen*: not already included in the given source page;
- *new*: providing new information (not outdated);
- *relevant*: worth the inclusion in the source page.

In our previous work, we analyzed the link between Wikipedia and the Question Answering task from another point of view: using Wikipedia in order to help finding answers [2]. Although the task is completely different from the *WiQA*, we achieved some experience in extracting useful informations from Wikipedia by means of a standard textual search engine such as Lucene².

¹ <http://ilps.science.uva.nl/WiQA/Task/index.html>

² <http://lucene.apache.org>

In this paper we describe the system we developed for the WiQA pilot task, that is mostly based on the Lucene search engine and our previous experience in mining knowledge from Wikipedia.

2 System Description

In Figure 1 we show the architecture of our system. A WiQA topic (i.e., a title of a source page to be expanded) is passed as a phrase (e.g. “Patriarch of Alexandria”) to the Lucene search module, which performs a standard search over the previously indexed collection of Wikipedia articles (the collections are the English and Spanish versions of the Ludovic Denoyer’s Wikipedia XML corpora [3]). Lucene returns a set of snippets ranked accordingly to the usual $tf \cdot idf$ weighting. The snippets at this points include those belonging to the source page, which are removed from the result set. Subsequently, the *ranking module* rearranges the snippets in the optimal ranking, that is the final result. The following section describes how the ranking is done.

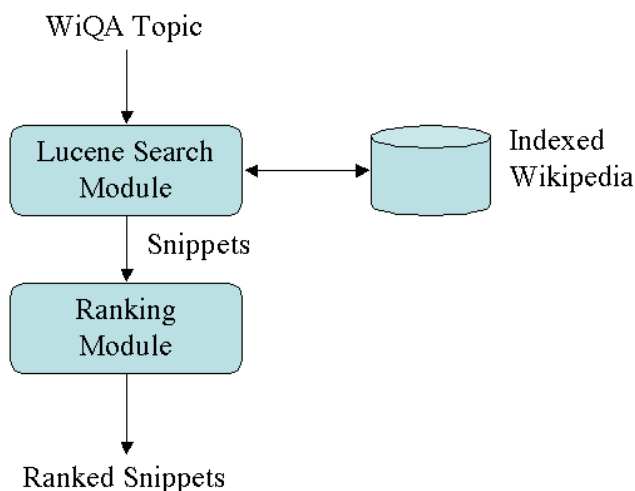


Fig. 1. Architecture of our WiQA system

3 Bag-of-Words Ranking

In our previous works on Wikipedia and QA we attempted to emulate a human use of Wikipedia by means of simple Lucene queries based on pages’ titles. The WiQA task presents almost the same characteristics, every topics being itself a page title. Our approach is straightforward and needs only to pass to Lucene a phrase query containing the source page title.

The user behaviors emulated by our system are the following:

1. The user searches for pages containing the title of the page he is willing to expand;
2. The user analyzes the snippets, discarding the ones being too similar to the source page.

In the first case, passing the title as phrase to Lucene is enough for most topics. We observed that some topics needed a better analysis of title contents, such as the 7th of the English monolingual test set: “*Minister of Health (Canada)*”, however these cases were not so many with respect to the total number of topics.

In the second case, the notion of *similarity* between a snippet and a page is the key for obtaining unseen (and relevant) snippets. Note that we decided to bound together relevance and the fact of not being already present in the page; we did not implement any method in order to determine whether the snippets contain outdated informations or not. In our system, the similarity between a snippet and the page is calculated by taking into account the number of terms they share. Therefore, we define the similarity $f_{sim}(p, s)$ between a page p and a snippet s as:

$$f_{sim}(p, s) = \frac{|p \cap s|}{|s|} \quad (1)$$

Where $|p \cap s|$ is the number of terms contained in both p and s , and $|s|$ is the number of terms contained in the snippet. This measure is used to rearrange the ranking of snippets by penalizing those being too similar to the source page. If w_l is the standard *tf · idf* weight returned by Lucene, then the final weight w of the snippet s with respect to page p is:

$$w(s, p) = w_l \cdot (1 - f_{sim}(p, s)) \quad (2)$$

The snippets are then ranked according to the values of w .

4 Results

The obtained results are shown in Table 4. The most important measure is the average yield, that is, the average number of “good” snippets per topic among top 10 snippets returned. The best average yield of the systems participating to the task was up to 3.38 for the English monolingual subtask [5]. We remark that our system ranked as the best one in Spanish, although the number of participants was smaller than for the English task.

The *Average yield* is calculated as the total number of important novel non-repeated snippets for all topics, divided by the number of topics. The *MRR* is the Mean Reciprocal Rank or the first important non-repeated snippet. The precision is calculated as the number of important novel non-repeated snippets, divided by the total number of snippets per topic. Our system returned always a number of snippets ≤ 10 .

Table 1. Results obtained by our system

Language	Average yield	MRR	Precision
English	2.6615	0.4831	0.2850
Spanish	1.8225	0.3661	0.2273

5 Conclusions and Further Work

Although we ranked as the best system in Spanish, the obtained results were worse than those obtained in English. We suppose this is due to the smaller size of the Spanish Wiki (117MB in contrast to 939MB): information is spread over a smaller number of pages, reducing the chance of obtaining valuable snippets. Our next steps in the WiQA task will be the inclusion of a similarity distance based on n-grams, resembling the one we already used for the JIRS Passage Retrieval engine [4]. We also hope to be able to participate in future cross-language tasks.

Acknowledgements

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work. This paper is a revised version of the work titled “A Naïve Bag-of-Words Approach to Wikipedia QA” included in the CLEF 2006 Working Notes.

References

1. Ahn, D., Jijkoun, V., Mishne, G., Mller, K., de Rijke, M., Schlobach, S.: Using wikipedia at the trec qa track. In: TREC 2004 Proceedings (2004)
2. Buscaldi, D., Rosso, P.: Mining knowledge from wikipedia for the question answering task. In: LREC 2006 Proceedings (July 2006)
3. Denoyer, L., Gallinari, P.: The Wikipedia XML Corpus. SIGIR Forum (2006)
4. Gómez, J., Montes, M., Sanchis, E., Rosso, P.: A passage retrieval system for multi-lingual question answering. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, Springer, Heidelberg (2005)
5. Jijkoun, V., de Rijke, M.: Overview of wiqa 2006. In: Peters, C. (ed.) Working notes of the CLEF 2006 Workshop, Alicante, Spain (2006)
6. Lita, L.V., Hunt, W.A., Nyberg, E.: Resource analysis for question answering. In: ACL 2004 Proceedings. Association of Computational Linguistics (July 2004)

University of Alicante at WiQA 2006

Antonio Toral Ruiz, Georgiana Pușcașu*, Lorenza Moreno Monteagudo,
Rubén Izquierdo Beviá, and Estela Saquete Boró

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{atoral,georgie,loren,ruben,stela}@dlsi.ua.es

Abstract. This paper presents the participation of University of Alicante at the WiQA pilot task organized as part of the CLEF 2006 campaign. For a given set of topics, this task presupposes the discovery of important novel information distributed across different Wikipedia entries. The approach we adopted for solving this task uses Information Retrieval, query expansion by feedback, novelty re-ranking, as well as temporal ordering. Our system has participated both in the Spanish and English monolingual tasks. For each of the two participations the results are promising because, by employing a language independent approach, we obtain scores above the average. Moreover, in the case of Spanish, our result is very close to the best achieved score. Apart from introducing our system, the present paper also provides an in-depth result analysis, and proposes future lines of research, as well as follow-up experiments.

Categories and Subject Descriptors: H.3[Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software;

General Terms: Measurement, Performance, Experimentation.

Keywords: Information Retrieval.

1 Introduction

Wikipedia¹ is a multi-lingual web-based, free content encyclopedia, continuously updated in a collaborative way. It may be seen as a paradigmatic example of a huge² and fast-growing source of written natural language.

Several inherent characteristics of this resource, such as its continuous growing nature, its general domain coverage, as well as its multilinguality, make Wikipedia a valuable resource for the Natural Language Processing (NLP) research field. The NLP community has only just lately become aware of this fact and started investing research effort in possible ways of exploiting Wikipedia within strategic areas such as Question Answering [2] or Knowledge Acquisition [8].

* Currently on research leave from University of Wolverhampton, United Kingdom.

¹ www.wikipedia.org

² By July 2006, the English version alone contains more than 1,250,000 entries.

WIQA³ is a pilot task at CLEF 2006⁴ exploiting the fact that, in Wikipedia, the distinction between author and reader has become blurred. The aim of the task is to discover how Information Retrieval and NLP techniques can be effectively used to help readers and authors of articles get access to information spread throughout Wikipedia rather than stored locally on a single page [4]. In a nutshell, WiQA is about collecting information about a certain topic not yet present on its page, thus avoiding data sparseness and unifying related content distributed among different entries. The motivation to launch this task lies in the already pointed out challenges that Wikipedia poses to the NLP community.

This paper is organized as follows. The following section presents a description of our approach and the developed system. Section 3 describes the experiments submitted to the WIQA task. Afterwards, in section 4, we present and comment the obtained results. Finally, in section 5, conclusions are drawn and future lines of research are pointed out.

2 System Description

Inspired by the Novelty and the QA tasks at TREC, the WiQA pilot task aims at recovering information not explicitly mentioned on a page, but distributed across the entire encyclopedia. The envisaged participating systems should help provide access to, author and edit Wikipedia's content. They should, mainly, be able to locate relevant and new sentences within the Wikipedia document collection, in response to a topic.

The WiQA task could be tackled from two perspectives, by employing either Information Retrieval methods, or Question Answering capabilities. Our approach embraces the IR strategy. Information Retrieval [3] is a Natural Language Processing application that, given a query and a document collection, returns a ranked list of relevant documents in response to the input query. IR usually comprises two stages. The first is a preprocessing phase which is carried out offline and consists of indexing the document collection. Its aim is to represent the documents in a way that makes it easier and more efficient to store and interrogate the collection. The second step, carried out online, consists of the actual retrieval of relevant documents as answer to an input query.

The architecture of our system is depicted in Figure 1. IR forms the core part of the system, being used to retrieve the documents relevant to the most meaningful terms in the topic document. For example, considering the topic *Alice Cooper*, we first extract the most meaningful terms in the supporting topic document. Then, in order to retrieve documents not only mentioning *Alice Cooper*, but also belonging to the domain defined by the extracted terms, we search the collection using these relevant terms.

For the above presented purposes, we have employed a probabilistic open source IR library called Xapian [1]. Besides the probabilistic search capability, in which the most relevant words are given increased weight, it allows boolean

³ <http://ilps.science.uva.nl/WiQA/>

⁴ <http://www.clef-campaign.org/>

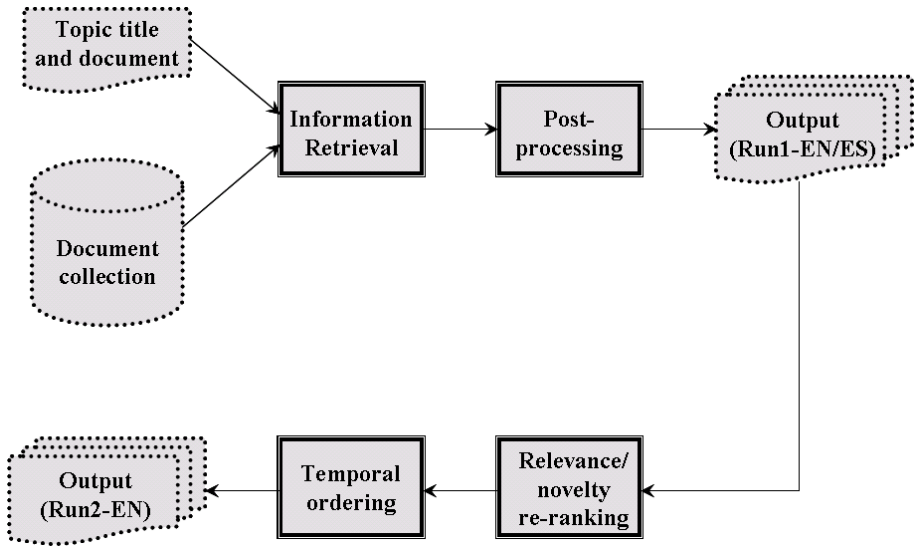


Fig. 1. System architecture

searches with operators that affect the query words, thus placing user-defined constraints on the search. These operators allow the user to specify, for example, that the desired terms occur in close proximity to each other. Another useful feature of Xapian is the possibility to receive feedback. By using this technique, Xapian can extract relevant terms for the query and carry out an expansion of it. The system first performs a basic retrieval and then gives the user the opportunity to select a set of documents considered relevant. At the next step, Xapian extracts from the selected documents relevant terms for query expansion. Finally, a second retrieval is performed by adding these terms to the query.

Due to the nature of WiQA pilot task the indexation has been performed at sentence level, that is sentences have been resembled to complete documents, therefore each indexed document is made up of only one sentence. This makes it straightforward to retrieve directly sentences in order to be compliant with the desired output of the system. In consequence, our system comprised a preprocessing phase that consisted of document sentence splitting and SGML tags removal. Finally, all the sentences contained in the document collection are indexed.

At the retrieval stage, Xapian has been configured using options and parameters to be presented in detail in the next section 3. Once the relevant sentences have been identified, they have been passed through a post-processing stage consisting of the following actions:

- Eliminate those sentences belonging to the topic supporting document
- Eliminate those sentences in documents linked from the query document (a link in the topic supporting document points to the sentence document). This elimination relied on the assumption that the information in the documents directly linked to the topic document was easily accessible through the links.

At this stage, a core set of sentences possibly relevant and important for the topic in question has already been delimited. In the case of the Spanish monolingual task, this set of sentences forms the system output, while, for the English task, they pass through subsequent processing stages, as described in the following.

Our English system proceeds by parsing the set of possibly relevant sentences, as well as the topic document sentences, with the Conexor's FDG Parser [7,5]. For the two sets of sentences and supporting document titles, the time expressions (TEs) are also identified and resolved using the temporal expression recogniser and normaliser previously developed by one of the authors [6]. The novelty of each retrieved sentence with respect to the sentences included in the Wikipedia topic document is then measured, in order to preserve the most relevant sentences to update the content of the Wikipage. Therefore, the sentences manifesting a high degree of similarity with the content of the topic document were characterised by very low scores.

The novelty degree of a retrieved sentence with respect to a sentence from the topic's Wikipage was considered to be a weighted measure revealing the percentage of novel named nouns (all uppercase nouns situated in the middle of the sentence), non-matching temporal expressions, novel nouns and verbs included in the former sentence, but not in the latter one.

The retrieved sentences are then ranked with respect to their novelty score, and passed on to a temporal ordering module. The temporal ordering module labels each retrieved sentence with the first TE occurring in the sentence, or, if no TE is present in the sentence, with the first TE of the title, or, if still no TE is found, with no label. Afterwards, the labelled sentences are interchanged so that their new order reflects their temporal order. The unlabelled sentences thus preserve their rank reflecting their novelty.

3 Description of Submitted Runs

Our WiQA submission includes one run for the Spanish monolingual task and two runs for the English task. The Spanish run and the first English run have been obtained with the same methodology, but applied to the language specific text collections. The second English run was obtained by employing more sophisticated NLP techniques to rank the set of retrieved sentences according to novelty, and to order them chronologically.

The first English run, as well as the Spanish run, both employ only the Information Retrieval capabilities of our system. Xapian firstly performs a search for the topic title constrained by the NEAR operator. The NEAR operator helps in locating the topic title words situated in any order within a short distance from each other. The feedback characteristic of Xapian is then employed to extract the important terms from the defined set of relevant documents (we classify as relevant the first 50 retrieved documents). Then the query is expanded with the identified relevant terms and a second retrieval is performed, this time constrained by the PHRASE operator. The PHRASE operator identifies only the

sentences where the group of words defining the topic occur together and in the same order as in the query. After post-processing the retrieved sentences by filtering out the ones that occur either in the topic document or in any document linked from the topic document, we preserve only the first 10 resulted sentences and return them as result of the first English run and of the Spanish run respectively.

Our second English run performs, apart from Information Retrieval, a novelty-based ranking, followed by a temporal ordering stage. Information Retrieval is employed in the same manner and with the same specifications as in the case of the first run. The retrieved sentences together with the topic sentences are parsed with the morpho-syntactic parser and with the temporal expression identifier/normaliser described above. A measure of novelty is then computed for each retrieved sentence with respect to all sentences from the topic document in turn, and the minimum score obtained will represent its degree of novelty with respect to the entire topic document. A novelty ranking of the retrieved sentences is then produced and passed on to a temporal ordering module. The temporal ordering module produces a new order of the sentences that reflects their succession on the temporal axis.

4 Results and Discussion

In this section we present and comment on the obtained results. As already stated, we have submitted three runs for WiQA 2006. Two runs represent solutions for the English monolingual task (one using IR only and the other one employing extra re-ranking and temporal ordering capabilities). The third run employs only IR and corresponds to the Spanish monolingual task.

The following two tables summarize the results for the monolingual English (Table 1) and for the monolingual Spanish (Table 2) tasks.

For each run, three different measures are provided (average yield, MRR and precision), all measured for the top 10 snippets returned. The average yield represents the average number of supported & novel & non-repetitive & important snippets retrieved. The MRR (Mean Reciprocal Rank) score refers to the first supported & novel & non-repetitive & important snippet returned. The precision was calculated as the percentage of supported & novel & non-repetitive

Table 1. Results for the English monolingual task

Run ID	Average Yield	MRR	Precision
1	2.98	0.53	0.33
2	2.63	0.52	0.32
MIN	1.52	0.30	0.20
MED	2.46	0.52	0.32
MAX	3.38	0.59	0.37

Table 2. Results for the Spanish monolingual task

Run ID	Average Yield	MRR	Precision
1	1.76	0.36	0.22
MIN	1.02	0.29	0.16
MED	1.06	0.30	0.22
MAX	1.82	0.37	0.27

& important snippets encountered among the submitted snippets. Apart from the results achieved by our runs, the two tables also include the minimum, median and maximum scores obtained for the tasks at hand. We are therefore able to evaluate and compare the performance of our system with respect to other participants.

The results show that our approach using feedback-driven IR has obtained results above the median value, both for English (Table 1) and for Spanish (Table 2). The fact that we score considerably better for English than for Spanish, though using the same approach, (average yield 2.98 EN vs. 1.76 ES, MRR 0.53 EN vs. 0.36 ES and precision 0.33 EN vs. 0.22 ES) might be due to the different size of the Spanish Wikipedia in comparison with the English version. The English version being notably larger than the Spanish version, there might be many more topic relevant text snippets spread across its entries.

Regarding the run which uses, apart from IR, novelty re-ranking plus temporal ordering, it has been submitted only for English, as it makes use of language dependent tools. The obtained results have been slightly worse than the ones for the first run. Our expectations were that these post-processing stages would at least bring a slight improvement to the results given by the IR engine alone. However, the performance has decreased. An in-depth analysis is needed to find out the causes of this unexpected behavior. Our opinion is that further investigation is required to improve or discover a more appropriate formula to be employed for measuring the novelty degree of a retrieved snippet. Besides, temporal processing should probably be employed at a point when WiQA systems are more mature and the input snippets are more reliable.

5 Conclusions

This paper presents our approach and participation in the WiQA 2006 competition. We propose feedback-driven IR with query expansion in order to retrieve, in response to given Wikipedia entries, relevant information scattered throughout the entire encyclopedia. Moreover, we have introduced a post-processing stage consisting of novelty ranking and temporal ordering. Our system participated both in the Spanish and English monolingual tasks.

When compared to the other systems presented in this competition, we have obtained good results that situate us above the medium score and quite close to the best result in the case of Spanish. Therefore, we conclude that the proposed

approach is appropriate for the WiQA task and we plan to find ways of improving the system's performance.

Several future work directions emerge naturally from a first look and shallow analysis of the results. Firstly, we would like to carry out an in-depth study of the effects induced by applying novelty ranking and temporal ordering, as the results obtained have not been those expected. Secondly, we aim at further investigating this topic departing from our feedback-driven Information Retrieval approach.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICYT number TIC2003-07158-C04-01.

References

1. Xapian: an Open Source Probabilistic IR library (visited 2006-06-01), On line www.xapian.org
2. Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., Schlobach, S.: Using wikipedia at the trec qa track. In: The University of Amsterdam at QA@CLEF 2004 (2005)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval (1999)
4. Jijkoun, V., de Rijke, M.: A Pilot for Evaluating Exploratory Question Answering. In: Proceedings SIGIR 2006 workshop on Evaluating Exploratory Search Systems (EESS) (2006)
5. Moreno-Montegudo, L., Suarez, A.: Una Propuesta de Infraestructura para el Procesamiento del Lenguaje Natural. In: Proceedings of SEPLN 2005 (2005)
6. Puscasu, G.: A Framework for Temporal Resolution. In: Proceedings of the 4th Conference on Language Resources and Evaluation (LREC2004) (2004)
7. Tapanainen, P., Jaervinen, T.: A Non-Projective Dependency Parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, ACL, (1997)
8. Toral, A., Muñoz, R.: A proposal to automatically build and maintain gazetteers for named entity recognition using wikipedia. In: Workshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy (2006)

A High Precision Information Retrieval Method for WiQA

Constantin Orăsan and Georgiana Puşcaşu

Research Group in Computational Linguistics
University of Wolverhampton, UK
{C.Orasan, georgie}@wlv.ac.uk

Abstract. This paper presents Wolverhampton University's participation in the WiQA competition. The method chosen for this task combines a high precision, but low recall information retrieval approach with a greedy sentence ranking algorithm. The high precision retrieval is ensured by querying the search engine with the exact topic, in this way obtaining only sentences which contain the topic. In one of the runs, the set of retrieved sentences is expanded using coreferential relations between sentences. The greedy algorithm used for ranking selects one sentence at a time, always the one which adds most information to the set of sentences without repeating the existing information too much. The evaluation revealed that it achieves a performance similar to other systems participating in the competition and that the run which uses coreference obtains the highest MRR score among all the participants.

1 Introduction

The Research Group in Computational Linguistics at the University of Wolverhampton has participated in the English task of the WiQA pilot task [2] with a system that employs high precision retrieval and shallow rules to rank candidate passages. Given that this was the first time such a task was proposed, our emphasis was less on fine tuning the system, and more on exploring the issues posed by the task and developing a complete system that can participate in the competition. The main goal of a system participating in the WiQA competition is the extraction of sentences from Wikipedia which contain the most relevant and novel information with respect to the content of the Wikipedia page describing the topic. Therefore, many important issues must be taken into consideration for a WiQA system development. First, it is necessary to identify the topic relevant sentences. Then, in the case of our system, the process of selecting and ranking the most important and novel sentences adopts a greedy strategy that adds to the topic document sentences from the retrieved set that bring the highest information gain. The relevant sentences are then presented as the system's output in the order they were selected.

This paper presents the outline of our system and the results of its evaluation. The structure of the paper is as follows: Section 2 describes the architecture of the system. The three runs submitted in the competition are described in Section 3. The evaluation results are presented and discussed in Section 4. In the last section, conclusions are drawn and future system enhancements are considered.

2 System Description

WiQA, as a newly introduced pilot task, cannot be characterised by typical approaches for system architecture. Still, three perspectives were foreseen for addressing it: based on Information Retrieval (IR), based on Question Answering (QA) or based on multi-document summarisation capabilities. In the IR perspective, a retrieval engine is the main source of candidate snippets which are then filtered and scored by a set of linguistic processors. The advantage of this method is that it does not rely on too much information about the topics to be processed. In contrast, in the QA-based approach the system knows what kind of information needs to be presented about a topic and employs a QA system to retrieve this information. For the WiQA task, whenever the answer is not contained in the topic page, the sentence containing the answer is returned as candidate snippet. Still, this template-based question answering process is limited to certain factoid questions, therefore losing access to possibly important information that does not fit any template. This approach can prove effective because with a small fraction of question types one could account for relevant information a user is interested in knowing, and the retrieved passages would thus have higher chances of being considered relevant, non-repetitive, and innovative in the case of questions that are not supported by the topic document. The approach based on multi-document summarisation treats Wikipedia as a collection of documents which needs to be summarised, and produces a multi-document summary focused on the topic. In some aspects this approach is similar to the one based on information retrieval because it also requires a retrieval engine, but the techniques employed to extract snippets can differ.

After considering the advantages and disadvantages, as well as the resources we had available to approach the task, we decided to develop an IR-based system for WiQA. Its architecture is depicted in Figure 1 and consists of four core components:

1. high accuracy retrieval engine
2. snippet pre-processor
3. snippet ranking module
4. snippet post-processor

In the remainder of this section each of these components is described in detail.

2.1 High Accuracy Retrieval Engine

The role of this component is to retrieve snippets on the selected topic from Wikipedia and to filter out snippets according to some predefined rules. The information retrieval engine used in our system is Lucene [3]. Given that the topics referred to named entities such as persons, locations or organisations, it was decided not to process the topic in any way and to query the retrieval engine for the exact string identifying the topic, in this way making sure that we have a high precision retrieval (i.e. all the retrieved snippets contain the string identified by the topic). Given that the majority of topics were unambiguous, most of the retrieved snippets were related to the topic.

The drawbacks of this method are that for several topics we were unable to obtain any candidates due to the way the topic was formulated, and that sentences which

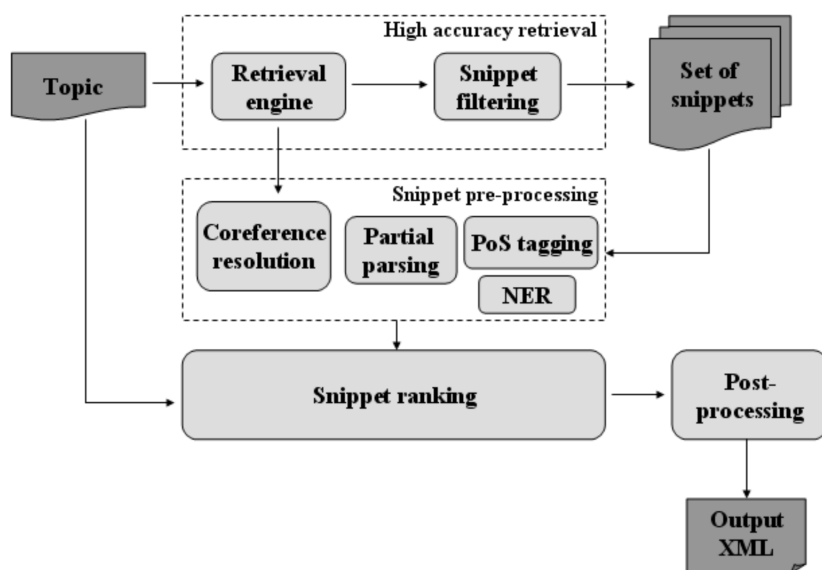


Fig. 1. System architecture

refer to the topic but do not contain the exact string identifying the topic cannot be extracted in this manner. The former problem appeared for topics such as *Ministry of Defence (Singapore)*, where the string *Singapore* was provided to disambiguate between different *ministries of defence*. In this case, when Lucene was queried with the string *Ministry of Defence (Singapore)*, no result was returned.

Not all the snippets retrieved by the engine can and should be used. The **Snippet filtering module** filters out snippets which are not appropriate for processing. First of all, the snippets extracted from the topic document have to be filtered out because they should not be considered for ranking, but are used to determine the ranking of the extracted snippets. Depending on the approach taken, snippets from documents directly connected to the topic document can be ignored because it can be argued that they contain information which is easily accessible to the reader of a document. Some of the runs we submitted used this filter.

In addition to the sentences extracted by our system, it was noticed that there are sentences which refer to the topic, but do not contain the actual string which identifies the topic. As a result, they cannot be directly extracted by the retrieval engine. Because in the majority of cases, these sentences refer to the topic using a referential expression, a coreference resolution module was added at the **Snippet pre-processing** stage and used in one of our runs.

The formatting of Wikipedia also determined us to implement some further filters. In some cases the snippets to be extracted are parts of a list and therefore it needs to be decided whether to ignore them, or whether to extract only the item in the list which contains the topic. Our runs used both types of filtering.

2.2 Snippet Pre-processing

At this stage, both the snippets extracted by the retrieval component and the sentences contained in the Wikipedia page describing the topic are annotated with linguistic information. Conexor's FDG Parser [4] is used to produce morpho-syntactic information and partial parsing trees for snippets. On the basis of this information, we identify noun phrases (NPs) and verb phrases (VPs), and to convert the snippets into logical forms. We distinguish between several predicates connecting nouns / verbs and corresponding to different syntactic relations, such as the relation between a verb and the head of its subject NP, between a verb and its object NP head, between a verb and its other complements and adjuncts, between two verbs and between two nouns. The derived logical forms have been only superficially employed at the snippet ranking stage, as more investigation is needed into ways of taking better advantage of the information they provide.

Given that many of the snippets contain named entities which can prove important at the ranking stage, GATE [1] was used to extract this information. It should be pointed out that the named entity recogniser provided by GATE has been used without any customisation and as a result it failed to recognise a large number of entities. In the future we plan to adapt the named entity recogniser in order to improve its recognition accuracy.

As already mentioned, not all the snippets relevant to the topic can be extracted by the search engine because they do not contain the topic string. In light of this, a rule based coreference resolver was used to identify sentences related to the topic. The rules implemented referred mainly to presence of a substring from the topic in a sentence following an extracted snippet. For example, in the following sentences:

- (a) *After returning home, Liang went on to study with **Kang Youwei**, a famous Chinese scholar and reformist who was teaching at Wanmu Caotang in Guangzhou.*
- (b) ***Kang**'s teachings about foreign affairs fueled Liang's interest in reforming China.*

using coreference resolution, it is possible to determine that *Kang* refers to *Kang Youwei*, and in addition to sentence (a), sentence (b) should be also extracted as candidate for the *Kang Youwei* topic. By using coreference resolution it is possible to expand the list of candidate snippets without introducing too much noise. In order to achieve this, documents which contain useful snippets are processed by the coreference resolver which identifies additional snippets.

2.3 Snippet Ranking

The **Snippet ranking** module uses information obtained at the pre-processing stage by the partial parser and NE recogniser in order to rank the retrieved snippets according to importance and novelty with respect to the topic document. The ranking algorithm is based on a greedy strategy which scores snippets by measuring how much extra information a snippet brings to a topic, and extracts those with the highest score. The algorithm is presented below:

1. Let $\mathbf{T}$ be the set of sentences extracted from the Wikipedia page corresponding to the topic, $\mathbf{R}$ the set of sentences retrieved from other documents, and $\mathbf{C}$ the current set of relevant sentences. $\mathbf{C}$ is initially empty ($\mathbf{C} = \emptyset$).
2. For each retrieved sentence $s \in \mathbf{R}$, calculate the information gain obtained by adding it to the set of topic sentences $\mathbf{T}$ unified with the current set of relevant sentences $\mathbf{C}$ (how much information do we gain having $\mathbf{T} \cup \mathbf{C} \cup \{s\}$ when compared to $\mathbf{T} \cup \mathbf{C}$).
3. Choose the sentence $s_{max} \in \mathbf{R}$ with the highest information gain score, add it to the set of relevant sentences $\mathbf{C}$, and, at the same time, eliminate it from the set of retrieved sentences ($\mathbf{C} = \mathbf{C} \cup \{s_{max}\}$, $\mathbf{R} = \mathbf{R} \setminus \{s_{max}\}$).
4. Continue selecting sentences that maximise the information gain score by repeating steps 2 and 3 until the maximum information gain value falls below a certain threshold, or until there is no candidate sentence left in $\mathbf{R}$.

The relevant/novel sentences resulted by applying this algorithm are presented, in the order they were added to set $\mathbf{C}$, as the output of the **Snippet ranking** module.

2.4 Post-processing

The initial purpose of the **Snippet post-processor** was to take the results from the **Snippet ranking** module and produce the XML file to be submitted for evaluation. After a script was written to produce this output, it was noticed that it did not produce the necessary output due to errors introduced by the FDG Parser. Because the input to the FDG tagger contained all kind of unusual characters, in a large number of cases snippets were wrongly split into several sentences. In other cases snippets were merged together even though they were on different lines due to the fact that they did not contain the final full stop. These errors were identified too late to be able to make any changes to the pre-processing module. In light of this, in order to produce the XML output required for the evaluation, a simple overlap measure between the ranked set of relevant sentences, as they were segmented by FDG, and the original snippets has been implemented in order to determine the source document and the snippet IDs, as required in the XML document. Due to the simplicity of this measure, a number of errors have been introduced in the output.

3 Description of Submitted Runs

We have submitted three runs referred to in the following as *one*, *one-old* and *two*. All three use the same retrieval engine, named entity recogniser, part-of-speech tagger, partial parser and post-processor. The runs *one* and *one-old* have in common the same set of extracted snippets, but the method in which they are ranked differs. For run *two* the set of snippets is different, but the same ranking method as that in run *one* is used.

For runs *one* and *one-old* the set of candidate snippets consists of only snippets which contain the topic string, without employing coreference resolution to expand the set. In addition, the sentence candidate set includes items from lists which contain the topic string, but not the whole lists. Snippets from documents directly linked to the topic document were discarded. In contrast, run *two* employs coreference resolution to

extend the set of candidates, does not filter out snippets from documents linked to the topic document, and does not include snippets from lists. As a result of this, the snippets contained less noise, but their number was also reduced.

The ranking method of run *one-old* differs from the ranking method of runs *one* and *two* in the formula employed for computing the information gain obtained by adding a sentence to a set of sentences. The run *one-old* uses formula (F1), while the runs *one* and *two* use a modified version of (F1) by increasing / decreasing the information gain score according to the following criteria:

- the score is penalised if no verbs are present in the snippet, or if the number of NEs or the number of occurring years (any 4-digit number greater than 1000 and smaller than 2020 is considered a possible year) are above 20 (if any of these cases apply, there is a high probability the snippet is a list or a list element);
- a small penalty applies in the case the topic is preceded by a preposition;
- the sentence is given a bonus if the topic is its subject;
- the sentence is penalized proportionally with the number of third person pronouns it contains.

$$(F1) \text{ IG}(s, \text{Set}) = (\mathbf{P}(t|\text{Set}) * \log \mathbf{P}(t|\text{Set}) - \mathbf{P}(t|\text{Set} \cup \{s\}) * \log \mathbf{P}(t|\text{Set} \cup \{s\}))$$

In the above formula, *s* and *Set* are the sentence and the sentence set we compute the information gain score for, *t* is a token which ranges over the set of NEs, nouns, verbs and logical forms present in the sentence, and $\mathbf{P}(t|\text{Set})$ is the probability of the token *t* to appear in the *Set*.

4 Results and Analysis

As one of the purposes of the WiQA pilot task was to experiment with different measures for evaluating systems' performance, several scores are used at the assessment stage. The evaluation measures used were: *yield*, *average yield per topic*, *MRR* and *precision* at top 10 snippets and were proposed by the organisers of the task [2]. Table 1 presents our results in the evaluation as well as the overall results obtained by all the participants. As our system targeted high precision, at the cost of low recall, by retrieving only those snippets that almost certainly referred to the query topic, we did not expect high *yield* and *average yield* scores. As one can notice in Table 1, our *yield* and *average yield* scores are close to the median value achieved by all WiQA participants. One of our runs (run *two*) established the maximum value for the MRR measure. The low precision can be explained by the fact that our runs contains a large number of repetitions introduced by the **Post-processing module**. If the repeated entries were discarded from the submission, the precision would have been 0.35, 0.34 and 0.35 for runs *one-old*, *one* and *two* respectively.

Comparison between runs reveals some interesting results. The weighting method used for runs *one* and *two* lead to better values for MRR, but the retrieval module used for runs *one-old* and *one* ensure better precision. The inclusion of coreference resolution and of snippets from pages directly linked to the topic page leads to the best value for MRR. In light of this, it would be interesting to develop a method which integrates

Table 1. Evaluation results

Run	Yield	Average yield per topic	MRR	Precision
One-old	142	2.32	0.57	0.30
One	135	2.21	0.58	0.28
Two	135	2.21	0.59	0.25
Min	58	1.52	0.30	0.20
Max	220	3.38	0.59	0.36
Median	158	2.46	0.52	0.32

them into the retrieval method used in runs *one* and *one-old*. Investigation of the results obtained in run *two* reveals that coreference resolution proposes a large number of very good snippets for ranking, but very few of them are present in the output of the system. In light of this, the ranking method will need to be changed to take into account the fact that a snippet was extracted by the coreference resolver and as a result it may not have the topic string or a large number of named entities.

Coreference resolution should be applied not only to sentences which follow a sentence including the topic. It should be applied also to sentences which contain the topic. For example in the sentence *In 2001 he took over the role of Shylock in the Stratford Festival of Canada production of The Merchant of Venice after Al Waxman, who was originally scheduled to play the part, died.* it is not possible to know which is the antecedent of *he* without a coreference resolver, and therefore the snippet will be marked as unimportant.

Comparison between the results corresponding to the three different runs revealed inconsistencies in the evaluation process. In this analysis, we first compared automatically the first occurrence of each snippet to identify whether they have been annotated identically in the two evaluation files. Only the first occurrence has been considered due to the fact that a second occurrence is repetitive and consequently receives a different annotation. An inconsistency is considered to be any difference in the values of the attributes corresponding to the first occurrence of a snippet in the two runs for which the comparison is made. This analysis revealed:

- * 14 inconsistencies between runs *one* and *one-old*
- * 15 inconsistencies between runs *one* and *two*
- * 11 inconsistencies between runs *one-old* and *two*

The next stage of the analysis was performed only for the runs *one* and *one-old* and involved manually checking the automatically extracted inconsistencies. This stage has revealed that, out of 14 automatically extracted inconsistencies, there are 10 true inconsistencies, that is snippets that should have received the same annotation. For example, the snippet

- (1) *Police cordoned off several areas in the capital of Mal, around the National Security Services, Shaheed Ali Building (Police Headquarters), Republic Square, People's Majlis and Maldivian Democratic Party Headquarters.*

is marked as supported, important, novel and not repeated in run *one-old*, whereas in run *one* it is annotated as a repetitive snippet, even though it is the first snippet presented for topic 47 (*Maldivian Democratic Party*).

There were also four cases for which the annotation provided by the evaluators is correct and the inconsistency is justified. For example in the case of topic 44 (*Al-Arabiya*) the information contained in a supported, important, novel and not repeated snippet of run *one-old* (2) was presented in another form by a previous snippet in run *one* (3), therefore being correctly tagged as repetitive.

- (2) *Through the first three months of 2004, a number of attacks on journalists in the West Bank and Gaza Strip have been blamed on the Brigades as well, including the attack on the Arab television station Al-Arabiya's West Bank offices by masked men self-identifying as members of the Brigades.*
- (3) *Through the first three months of 2004, a number of attacks on journalists in the West Bank and Gaza Strip have been blamed on the Al-Aqsa Martyrs' Brigades, most clearly the attack on the Arab television station Al-Arabiya's West Bank offices by masked men self-identifying as members of the Brigades.*

These inconsistencies in the evaluation results are not a surprise given that the measures used to assess the snippets are subjective.

5 Conclusions

This paper presented Wolverhampton University's submission to WiQA task. The method chosen for this task combines a high precision, but low recall information retrieval approach with a greedy algorithm to rank the retrieved sentences. The results of the evaluation revealed that it achieves a performance similar to other systems participating in the competition and that one of the submitted runs scores the highest MRR score among all the participants.

System analysis pinpointed several possible improvements for the future. As noticed in the evaluation, better precision can be obtained by filtering out duplicates at the post-processing stage. More attention will need to be paid to the input of the pre-processing stage so the sentences are no longer wrongly identified. As mentioned previously, in order to compute the relevance of each sentence, the sentences are converted into a logical form. The predicates used for this representation are currently little used, but we believe that the ranking can be improved by using them.

References

1. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (2002)
2. Jijkounand, V., de Rijke, M.: Overview of WiQA 2006. In: Proceedings of CLEF2006 (2006)
3. Lucene: Lucene. <http://lucene.apache.org/>
4. Tapanainen, P., Jaervinen, T.: A Non-Projective Dependency Parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, ACL (1997)

QoIA: Fostering Collaboration Within QA

Diana Santos and Luís Costa

Linguatca, Oslo node and SINTEF ICT, Norway
Diana.Santos@sintef.no, Luis.Costa@sintef.no

Abstract. In this paper we suggest a QA pilot task, dubbed QoIA, whose joint rationale is allow for collaboration among systems, increase multilinguality and multcollection use, and investigate ways of dealing with different strengths and weaknesses of a population of QA systems. We claim that merging answers, weighting answers, choosing among contradictory answers or generating composite answers, and verifying and validating information, by posing related questions, should be part and parcel of the question answering process. The paper motivates these ideas and suggests a way to foster research in these areas by deploying QA systems as Web services.

1 Motivation

There were many reasons that led us to propose QoIA in the CLEF 2006 workshop in Alicante [18], which we would like to expand and further motivate here. Some are related to CLEF’s ultimate aims as we understand them: namely advance the field of crosslingual IR systems and provide a forum for researchers and developers to cross the cultural barrier of dealing only with their own language, especially in a European context. Even though we believe that it is rewarding in itself to be together with QA researchers from some 40 groups who together process 8 languages, we are convinced that a lot can still be done to improve cross-fertilization of the different approaches and expertise that has come together (for example, hardly any paper in the QA section mentions other QA papers in CLEF – especially not those dealing with other languages).

Other goals are related to the QA field in itself and our attempt to advance it, inspired by the suggestions in Burger et al.’s roadmap [5], and in Maybury [15] and Strzalkowski and Harabagiu’s [20] recent books (both of them, however, devote very little attention to crosslingual or multilingual issues). CLEF was the best place to do this because of the many collections in different languages it provides, and because it includes Portuguese. (Linguatca’s aim is to significantly improve the state of the art of Portuguese processing).

1.1 Going Beyond the “1-1 Model”

Question answering can be seen from two complementary perspectives, to which we refer to as the information retrieval (IR) and the artificial intelligence (AI) ones, inspired by Wilks’s analysis [24]:

- QA is more user-friendly information retrieval, for less computer-literate users;
- QA is (one of) the best application(s) to test a system's understanding.

Not surprisingly, depending on one's view on the ultimate meaning of QA (what is QA really about), efforts by QA developers and researchers are devoted to different endeavours:

On the one hand, the IR tradition takes it for granted that QA is about mas-saging the documents that are returned by IR, so that QA can be approximated into a translation of questions into query terms, with subsequent choice of the appropriate passage (or part of it) for the answer. The TREC setup [21] and Brill et al. [4] are good illustrations of such an approach; the lexical chasm model of Berger et al. [3] is probably the cleverest way of solving QA as IR.

On the other hand, the AI view of QA as testing the understanding of a system leads researchers to elaborate strategies of encoding knowledge and reasoning about it. Research under this paradigm is invariably concerned with understanding why the question was posed and what is expected of the answer (with the standard reference dating back to Lehnert [12]).¹

One can make here the analogy with, in the fields of summarization and speech synthesis, extractive or concatenative paradigms as opposed to generation or synthesized ones: the first approach seeks primarily to *find* something that is already there, while the second *creates* something from the raw material.

For both approaches, however, there is one common difficulty: a question-answer pair is basically describable as a M,N-O,P relationship between queries and answers: There are many ways (M) to pose the same question, and there are many questions (N) that get the same answer. Plus: there are many correct ways to give the same answer (O), and many correct answers that can be given to the same question (P). So, one essential problem in open-domain QA is, given a particular question, to provide or choose one particular answer (or an acceptable combination of several ones). Taking one step further, [11] even argue for an approach of considering different candidate answers as potential allies, instead of competitors.

Technically, this means that, no matter their specific perspective for the solution, (a) dealing with alternative answers, (b) dealing with alternative ways of posing questions, (c) ranking different answers and (d) combining them, are tasks relevant to every QA system.

Now, one way to handle a multitude of related questions and answers is invoking a population of automatic question answerers.² This way, one can not only gather several different (or similar) answers to a same question, but also pose similar or related questions, as suggested e.g. Wu and Strzalkowski [25]. This is

¹ Although this is not necessarily the only AI/NLP origin of QA: Dalmas and Webber [11], for instance, refer both to the natural interaction with databases and to the reading comprehension task as precursors of present-day "QA over unstructured data".

² A similar approach, invoking software agents, was first implemented and proposed by Chu-Carroll et al. [8], who report very positive results.

the main rationale for the QoLA setup, which, as a by-product, also aims to stimulate the collaboration among different approaches that use different resources in different languages and search in different collections.

1.2 Real Answers Are Usually Not Found in One Place Only

In addition to the general motivation above (the M,N-O,P model), most answers to a real question are often crucially dependent on the user context and intentions. So, we believe that emphasis should be put to harvesting sets of possible answers and defining methods to provide useful and informative answers instead of betting on the one-answer only model.³

In fact, as early as TREC 99 [21], with the apparently unequivocal question *Where is the Taj Mahal?*, it became clear that correctness of an answer could not be decided beforehand, and that much more useful than an answer like *Agra* would either be some tourist information on how to get there (e.g., *half an hour by train from New Delhi*) or the address of the Taj Mahal casino in New Jersey. Not only proper names can have many referents, though: even question words such as *where* in *where was X stabbed?* can be interpreted in radically different ways – and thus the equally relevant answers *in the neck* and *in Cairo* – as well as common concepts such as *football*: which kind? [17]. Also *where*-questions depend crucially on the context, so that they can even sidetrack a general semantic type-checking approach, as Schlobach et al.’s [19] discussion of *where is hepatitis located?* shows.

Furthermore, apparently neutral questions depend on both the answerer and questioner’s standpoints, and it is arguably better to provide more than one standpoint and leave the questioner to choose. For a “simple” question such as *Who was Freud?* we can get from the Web both *Freud was a terrible sadist who had the mood of a hysterical woman* and *Freud was a born leader with a deep knowledge of humanity and unerring judgment*. It is probably more interesting for most people interested in Freud to get a fair idea of the reactions to him than have an automatic system perform a choice for us.

In 2006 a substantial improvement was done to the QA@CLEF setup by accepting, not only a set of justification snippets per answer but a set of possible (ten) answers for each question as well. However, this new feature was apparently not used by most systems, which continued to send one answer only, probably because of the way they are conceived.⁴ QoLA would provide systems with the possibility to experiment with any number of answers, as well as with strategies to concatenate and/or decide on the final set. Instead of the quest for the only unique answer, QoLA was therefore thought to help experiment with merging answers, weighting answers, choosing among contradictory answers or generating

³ That this is a hopelessly wrong conceptual model is clearly displayed in [11]’s quantitative estimate of the number of distinct correct answers for TREC questions, which were originally conceived under the question-with-one-answer-only model.

⁴ To be fair, we should also mention that most questions provided by the QA@CLEF organizers were still of the one-correct-answer-only variety, so there was no real incentive at last year’s CLEF to return multiple answers.

composite answers, and verifying and validating information, by posing related questions. All these may be parts of the answer.

1.3 Going Beyond “One System at a Time” and Closer to the User

In the QA@CLEF evaluation setup, each system mainly competes with the others. Even though the CLEF spirit is crosslingual, most work so far [14] has been done on monolingual QA (each language has its own collection(s)), and at most bilingual retrieval: the joint use of more than one language collection has never been tried.

One of the possible drawbacks of QoLA is that, in a QoLA run, no single system would get the honour or responsibility of the answer. While this may diminish the individual zest of developers to provide the best individual system, we believe that, in a realistic setup, a user is not ultimately interested in how the answer has been arrived to, which system got it right, and which collections have been searched, provided s/he gets the source information (the justifications) and can confirm or discard the answer.

Ever since FAQFinder [6], there are growing repositories of previous answers and systems that mine them, so it is not realistic that future systems will depend only on their own proprietary bases created from scratch. On the contrary, harnessing the common knowledge and making use of many sources of information is key to the Web 2.0 perspective [16].

2 How to Operationalize Collaboration

The first challenge in QoLA was to provide a way to invoke the several participating QA systems, and to us the best solution would be that the participating QA systems were available as web services (WS). This was still not enough: the services had to be interoperable as well. If people defined their systems separately, they would most likely provide operations with different names, different parameters and parameter names, etc. Therefore, a task of the organization was to define a set of operations relevant for QoLA, which all providers of QA systems for QoLA had to obey.

These operations were specified using WSDL (web service description language), [23] which is a de facto standard for describing web services. A WSDL web service definition provides information about the names of operations and parameters, but not about the functionality of the service. There are several research initiatives which aim to define the semantics of web services (WSDL-S, OWL-S, WSMF, IRS) [2,7], but for this first edition of QoLA we decided to keep it simple and only cater for syntactic interoperability.

2.1 Web Service Population

The participating QA systems in QoLA should be available as web services according to the specification that we proposed (and which evolved considerably with the feedback from the potential participants).

At some point, Sven Hartrumpf from the University of Hagen suggested the inclusion of translation web services in addition to the QA web services. Even though we were expecting the participation of bilingual QA web services and had already stipulated that we would provide every question in all QoLA languages, we quickly agreed that separate translation Web services would make a lot of sense in the context of QoLA and CLEF in general.

It was therefore arranged that there would be two kinds of Web services in QoLA: QA services and machine translation (MT) services, provided by (special) participants called **QoLA providers**.

General participation in QoLA would be granted to anyone interested in invoking these services, and this kind of participants were dubbed **QoLA consumers**.

2.1.1 QA Web Service Definition

The QA web services should provide the following set of basic operations: *GetSupportedLanguagePairs*, *GetSupportedEncodings*, *GetQuotaInfo*, *GetResourcesUsed*, *GetModulesUsed*, *GetAccuracy* and *GetAnswer*:

GetSupportedLanguagePairs: This operation returns a list with the language pairs that the QA system can handle. For example, questions in Portuguese/answers in Spanish, questions in English/answers in Italian, etc.

GetSupportedEncodings: This operation returns a list with the character encodings accepted by the QA Web service (UTF-8, ISO-8859-1 or others).

GetQuotaInfo: It would be natural that QA web services had some mechanism to limit the number of requests per customer, not only for internal performance issues, but also to give a fair chance to all participants using the web service. *GetQuotaInfo* gives information about the available quota for a customer (how many questions can be asked to the QA web service for a certain period).

GetResourcesUsed: This operation returns a list of the resources employed by the QA system. This should include a list with names of ontologies, gazetteers, specific publicly available collections that are used by the system, such as WordNet 6.1, CHAVEanot, DutchWikipedia, TGN, etc. (These and other codes would have to be agreed for QoLA.) This list is important for a consumer to decide whether a same answer returned by different systems came from the same source or from different ones.⁵

GetModulesUsed: This operation returns a list with the internal modules used by the QA system (*CoreferenceResolution*, *DeixisResolution*, *InferenceEngine*, *AnswerValidation*, *NaturalLanguageGeneration*, *NamedEntityRecognizer*, *Parser* and other codes agreed among the participants and organizers of QoLA).

GetAccuracy: The idea behind this operation was to provide consumers with some information about the performance of a particular QA web service for

⁵ However, [8] make the important point that, even when two answering agents – in our case, two different systems – consult the same sources, they may arrive at an answer with significantly different processing strategies, which is a (partial) motivation for the next operation, *GetModulesUsed*.

a certain language pair. For example, the system could return accuracy for the 2006 questions in general, or for each kind of question. No consensus as yet has been reached as to obligatoriness and format of this operation (note that systems could have improved dramatically since 2006, or not have participated in that event).

GetAnswer: This operation receives a question, the language and character encoding of the question, the language of the desired answer and a customer code (that could be used to limit the number of allowed questions per customer). The answer would consist in a string with the answer itself, a confidence score, the character encoding of the answer, the list of resources used, the list of modules used, and a list of justifications. Each of the justifications contains a document ID, a snippet of the document and a confidence score for the justification. Note that even if the system were not able to answer, a NIL answer, with a confidence score, should always be provided.

2.1.2 MT Web Service Definition

The MT web services should provide the following set of basic operations: *GetSupportedLanguagePairs*, *GetSupportedEncodings*, *TranslateQuestion*, *TranslateNamedEntity* and *TranslateSnippet*:

GetSupportedLanguagePairs, GetSupportedEncodings: These operations return lists with the language pairs and encodings that the MT system can handle.

TranslateQuestion, TranslateNamedEntity, TranslateSnippet: These three operations are provided because different techniques can be used to translate questions, named entities and text snippets. All of these operations receive the text to be translated, its language and character encoding and the target language for the translation. The result is the translation of the source text to the target language and the character encoding of the translation.

2.2 The Task

In order to make the least changes, we decided to use exactly the same evaluation setup that would be used (and therefore organized) for QA@CLEF 2007 main task. So the evaluation of the QoLA results would be done the very same way the other sets of answers were dealt with by CLEF QA assessors. The way participants used the combination of the several WS available would make all the difference, we hoped.

In order to allow the future consumers to experiment with the full QoLA concept, and considering that there might be a period where QoLA providers had not yet made their systems available, Bernardo Magnini suggested the creation of a set of fake services which might only answer previous questions, but which could be made available by the organization, provided previous QA@CLEF participants allowed us to use their previous runs (and most at once did).

2.3 Baselines and Runs

To participate in QoLA, a consumer participant would have to provide a logic, in the form of a program, for invoking the different services available.

In order to encourage participants to really try collaboration, the organization should provide a large variety of simple baselines, such as: find the quickest answer, random choice, majority choice, confidence score choice, intersection and union choices, etc., so that participants were “forced” to do something better.

Also, QoLA’s organization disallowed “selfish” runs, in the sense of runs centered in one’s own system. This means that fixed schemas invoking any system by name were forbidden – only general runs were allowed, to prevent strategies like “let me first invoke my system, then the best of last year’s CLEF...”: No system should be given advantage that was not based on its own current behaviour.

3 Expected Results

(Consumer) participants were expected to experiment in QoLA with at least the following issues:

- validation of previous answers in other collections (possibly with yes/no questions), or with different but synonymous formulations (see Clarke et al. [9] and Magnini et al. [13] on validation approaches and exploiting redundancy for this purpose);
- mining related questions and related information on the same subjects (as advocated in [11]);
- integrate in different ways related information (including both Chu-Carroll et al.’s [8] “answer resolution” and the “answer tiling” of Ahn et al. [1]);
- use redundancy in collections or across collections to rank answers (see i.a. Clifton et al. [10]);
- devise strategies to separate same answers from really different answers;
- try to get some grasp of getting different answers in different contexts;
- investigate different answer typing strategies (see Schlobach et al. [19]);
- create (and use) a set of canonical questions to pose to different systems in order to test their knowledge in specific areas, like the University of Bangor in TREC 2002 [10];
- investigate relations among related questions (following [25]).

We note that invoking QA systems as Web services is virtually unlimited – and access to this kind of power (without bothering other people, just using automated systems) allows one to augment knowledge by orders of magnitude, which could then be used intelligently. (In fact, our colleagues from Priberam suggested that a useful byproduct of the QoLA pilot was to use its setup to create large collections of validated questions, provided specific interfaces were also developed.)

Another potential impact of QoLA is the possible experimentation with the deployment (and use) of special translation web services, that might work differently for specific areas, and for QA in particular, which could be very relevant to crosslingual QA, and CLIR in general. For example, multilingual ontologies and/or gazetteers would come handy for the specific translation of questions, which often have named entities (proper names) which are difficult to deal with by current MT systems. CLEF seems to be the right place to investigate and deploy such specific “query” translation systems, while a web service developed around EuroWordNet [22] would be a sensible candidate for a specific translation provider.

Finally, another interesting experiment would be merging or making sense of information in different languages: if one could use a set of different QA as WS to discover that *London* and *Londres* are the same location (by getting for example the answer in English and in Portuguese about *What is the capital of England?*), or that *Malvinas* and *Falkland* also name a very same location despite completely different names, this might allow systems to answer much more fully questions about these places.

4 Concluding Remarks

We were unfortunately unable to provide the necessary infrastructure for QoLA for 2007, and are therefore limited to suggesting the idea to the wider QA community. Some preliminary results are, however, available through <http://www.linguateca.pt/QoLA/>: the existence of some QA systems deployed as Web services, the formal WSDL specification of a question answering WS, and general discussion about the whole pilot and how to make it concrete. We still hope QoLA will take place some day, not necessarily with us organizing it.

Acknowledgements. The work reported here has been (partially) funded by Fundação para a Ciência e Tecnologia (FCT) through project POSI/PLP/43931/2001, co-financed by POSI, by POSC project POSC/339/1.3/C/NAC, and by SINTEF ICT through project 90512140. We thank Luís Sarmento for providing a translator Web service definition and one baseline, and Sven Hartrumpf, Johannes Levelling, Bernardo Magnini, André Martins and Luís Sarmento for participating in the discussion to make the pilot more concrete. We thank Paulo Rocha and Andreas Limyr for comments on preliminary versions of this paper.

References

1. Ahn, D., Fissaha, S., Jijkoun, V., Müller, K., de Rijke, M., Tjong Kim Sang, E.: Towards a Multi-Stream Question Answering-As-XML-Retrieval Strategy. In: The Fourteenth Text Retrieval Conference (TREC 2005) (2006)
2. Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., Verma, K.: Web Service Semantics. In: WSDL-S Proceedings of the W3C Workshop on Frameworks for Semantics in Web Services, Innsbruck, Austria, June 9-10, 2005 (2005)

3. Berger, A., Caruana, R., Cohn, D., Freitag, D., Mittal, V.: Bridging the lexical chasm: statistical approaches to answer-finding. In: Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00), Athens, Greece, July 24 - 28, 2000, pp. 192–199. ACM Press, New York (2000)
4. Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A.: Data-Intensive Question Answering. In: Voorhees, E.M., Harman, D.K. (eds.) *Information Technology: The Tenth Text Retrieval Conference (TREC 2001)*. NIST Special Publication 500-250, pp. 393–400 (2002)
5. Burger, J., et al.: Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) (October 24, 2003) http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/molla/qa_roadmap.pdf
6. Burke, R.D., Hammond, K.J., Kulyukin, V.A., Lytinen, S.L., Tomuro, N., Schoenberg, S.: Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System. Technical Report. UMI Order Number: TR-97-05, University of Chicago (1997)
7. Cabral, L., Domingue, J., Motta, E., Payne, T., Hakimpour, F.: Approaches to Semantic Web Services: An Overview and Comparisons. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) *ESWS 2004*. LNCS, vol. 3053, pp. 225–239. Springer, Heidelberg (2004)
8. Chu-Carroll, J., Prager, J., Welty, C., Czuba, K., Ferrucci, D.: A Multi-Strategy and Multi-Source Approach to Question Answering. In: Proceedings of TREC 2002. NIST Special Publication 500-251, pp. 281–288 (2003)
9. Clarke, C.L.A., Cormack, G.V., Lynam, T.R.: Exploiting redundancy in question answering. In: Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '01), New Orleans, Louisiana, United States, pp. 358–365. ACM Press, New York (2001)
10. Clifton, T., Colquhoun, A., Teahan, W.: Bangor at TREC 2003: Q&A and Genomics Tracks. In: Voorhees, E.M., Buckland, L.P. (eds.) *The Twelfth Text Retrieval Conference (TREC 2003)*. NIST Special Publication: SP 500-255, pp. 600–611 (2004)
11. Dalmas, T., Webber, B.: Answer comparison in automated question answering. *Journal of Applied Logic* 5(1), 104–120 (2007)
12. Lehnert, W.G.: *The process of question answering*. Lawrence Erlbaum Associates, Hillsdale, NJ (1978)
13. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Is It the Right Answer? Exploiting Web Redundancy for Answer Validation. In: Proceedings of 40th Anniversary Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA, July 6–12, 2002, pp. 425–432 (2002)
14. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2006, Multilingual Question Answering Track. (In This volume)
15. Maybury, M.T. (ed.): *New directions in question answering*. AAAI Press/the MIT Press, Cambridge (2004)
16. O'Reilly, T.: What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software (09/30/2005), <http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
17. Santos, D., Cardoso, N.: Portuguese at CLEF. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 1007–1010. Springer, Heidelberg (2006)

18. Santos, D., Costa, L.: Where is future QA?: One pilot and two main proposals. Presentation at the CLEF 2006 Workshop (Alicante, 22 September 2006) (2006), <http://www.linguateca.pt/documentos/SantosCostaQACLEF2006.pdf>
19. Schlobach, S., Ahn, D., de Rijke, M., Jijkoun, V.: Data-driven Type Checking in Open Domain Question Answering. *Journal of Applied Logic* 5(1), 121–143 (2007)
20. Strzalkowski, T., Harabagiu, S. (eds.): *Advances in Open Domain Question Answering*. Springer, Heidelberg (2006)
21. Voorhees, E.M., Tice, D.M.: Building a Question Answering Test Collection. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, July 2000, pp. 200–207 (2000)
22. Vossen, P.(ed.): *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht (1998)
23. W3C Note.: *Web Services Description Language (WSDL) 1.1*. (2001), <http://www.w3.org/TR/wsdl>
24. Wilks, Y.A.: Unhappy Bedfellows: The Relationship of AI and IR. In: John, I.T. (ed.) *Charting a New Course: Natural Language Processing and Information Retrieval: Essays in Honour of Karen Spärck Jones*, pp. 255–282. Springer, Dordrecht (2005)
25. Wu, M., Strzalkowski, T.: Utilizing co-occurrence of answers in question answering. In: *Proceedings of the 21st international Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, Annual Meeting of the ACL. Association for Computational Linguistics*, Sydney, Australia, July 17 - 18, 2006, pp. 1169–1176. Morristown, NJ (2006)

Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks

Paul Clough¹, Michael Grubinger², Thomas Deselaers³,
Allan Hanbury⁴, and Henning Müller⁵

¹ Sheffield University, Sheffield, UK

² Victoria University, Melbourne, Australia

³ Computer Science Department, RWTH Aachen University, Germany

⁴ Vienna University of Technology, Vienna, Austria

⁵ University and Hospitals of Geneva, Switzerland
`p.d.clough@sheffield.ac.uk`

Abstract. This paper describes the general photographic retrieval and object annotation tasks of the ImageCLEF 2006 evaluation campaign. These tasks provided both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information systems for image retrieval and automatic image annotation. Both tasks offered something new for 2006 and attracted a large number of submissions: 12 groups participated in ImageCLEFphoto and 3 groups in the automatic annotation task. This paper summarises these two tasks including collections used in the benchmark, the tasks proposed, a summary of submissions from participating groups and the main findings.

1 The Photographic Retrieval Task: ImageCLEFphoto

The ImageCLEFphoto task provides the resources for the comparison of system performance in a laboratory-style setting. This kind of evaluation is system-centred and similar to the classic TREC¹ (Text REtrieval Conference [1]) ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but search topics are not known to the system in advance. Evaluation aims to compare algorithms and systems, and not assess aspects of user interaction (iCLEF addresses this). The specific goal of ImageCLEFphoto is: given a statement describing a user information need, find as many relevant images as possible from the given document collection (with the query in a language different from that used to describe the images). After three years of evaluation using the St. Andrews database [2], a new database was used in this year's task: the *IAPR TC-12 Benchmark* [3], created under Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR²). This collection differs from the

¹ <http://trec.nist.gov/>

² <http://www.iapr.org/>

St Andrews collection used in previous campaigns in two major ways: (1) it contains mainly colour photographs (the St Andrews collection was primarily black and white) and (2) it contains semi-structured captions in English *and* German (the St Andrews collection used only English).

1.1 Document Collection

The ImageCLEFphoto collection contains 20,000 photos taken from locations around the world, comprising a varying cross-section of still natural images on a variety of topics (Fig. 1 shows some examples). The majority of images have been provided by *viventura*³, an independent travel company organising adventure and language trips to South America. Travel guides accompanying the tourists maintain a daily online diary including photographs of the trips made and general pictures of each location. For example, pictures include accommodation, facilities, people and social projects. The remainder of the images have been collected by the second author over the past few years from personal experiences (e.g. holidays and events). The collection is publicly available for research purposes and unlike many existing photographic collections used to evaluate image retrieval systems, this collection is very general in content. The collection contains many different images of similar visual content, but varying illumination, viewing angle and background. This makes it a challenge for the successful application of techniques involving visual analysis.



Fig. 1. Sample images from the IAPR TC-12 collection

The content of the collection is varied (and realistic) and associated descriptive annotations have been carefully created and applied in a systematic manner (e.g. all fields contain values of a similar style and format) to all images. Each image in the collection has a corresponding semi-structured caption consisting of the following seven fields (similar to the previous St Andrews collection): (1) a unique identifier, (2) a title, (3) a free-text description of the semantic and visual contents of the image, (4) notes for additional information, (5) the provider of the photo and fields describing (6) where and (7) when the photo was taken. These fields exist in English and German, with a Spanish version currently being verified. Although consistent and careful annotations are typically not found in

³ <http://www.viventura.de>

practice, the goal of creating this resource was to provide a general-purpose collection which could be used for a variety of research purposes. For example, this year we decided to create a more realistic scenario for participants by releasing a version of the collection with a varying degree of annotation “completeness” (i.e. with different caption fields available for indexing and retrieval). For 2006, the collection contained the following levels of annotation: (a) 70% of the annotations contain title, description, notes, location and date; (b) 10% of the annotations contain title, location and date; (c) 10% of the annotations contain location and date; and (d) 10% of the images are not annotated (or have empty tags respectively).

1.2 Query Topics

Participants were given 60 topics representing typical search requests for this document collection. The topic creation process is an important aspect of a test collection as one must aim to create a balanced and representative set of information needs. Topics for ImageCLEFphoto were derived from analysing the log file⁴ from a web-based interface to the image collection which is used by employees and customers of *viventura*. Domain knowledge from the authors was also used. To provide an element of control over the topics, the final set given to participants were based on considering a number of parameters including: geographical constraint, “visualness” of the topic, an estimation of linguistic difficulty, degree of annotation completeness and the estimated number of relevant images. Topic creators aimed for a target set size of between approximately 20 to 100 relevant images and thus had to further modify some of the topics (broadening or narrowing the concepts).

For many of the topics, successful retrieval using text-based IR methods required the use of query analysis (e.g. expansion of query terms or logical inference). These reflected examples found in the log files, e.g. for the query “group pictures on a beach”, many of the annotations would not use the term “group” but rather terms such as “men” and “women” or the names of individuals. Similarly for the query “accommodation with swimming pool” (also from the log file), the query would result in limited effectiveness unless “accommodation” was expanded to terms such as “hotel” and “B&B”. Queries such as “images of typical Australian animals” required a higher level of inference and access to world knowledge (this query was not found in the log file but could be a feasible request by users of an image retrieval system).

Similar to previous analyses of search log files (see, e.g. [4]), we found many search requests to exhibit some kind of geographical constraint (e.g. specifying a location). Therefore, we created 24 topics with a geographic constraint (e.g. “tourist accommodation *near Lake Titicaca*”), 20 topics with a geographic feature or a permanent man-made object (e.g. “group standing in *salt pan*”) and 16 topics with no geography (e.g. “photos of female guides”). All topics were classified regarding how “visual” they were considered to be. An average rating

⁴ Log file from 1st February–15th April 2006 containing 980 unique queries.

between 1-5⁵ was obtained for each topic from three experts in the field of image analysis, and the retrieval score from a baseline content-based image retrieval (CBIR) system⁶. A total of 30 topics were classed as “semantic” (levels 1 and 2) for which visual approaches would be highly unlikely to improve results (e.g. “cathedrals in Ecuador”); 20 topics classified as “neutral” (level 3) for which visual approaches may or may not improve results (e.g. “group pictures on a beach”) and 10 were “visual” for which content-based approaches would be most likely to improve retrieval results (e.g. “sunset over water”). To consider topics from a linguistic viewpoint, a complexity measure was used to categorise topics according to their linguistic difficulty [5]. A total of 31 “easy” topics were selected (levels 1 and 2, e.g. “bird flying”), 25 “medium-hard” (level 3, e.g. “pictures taken on Ayers Rock”), and 4 “difficult” topics (levels 4 and 5, e.g. “tourist accomodation near Lake Titicaca”). Various aspects of text retrieval on a more semantic level were considered too, concentrating on vocabulary mismatches, general versus specific concepts, ambiguous terms and use of abbreviations.

Each original topic comprised a title (a short sentence or phrase describing the search request in a few words), and a narrative (a description of what constitutes a relevant or non-relevant image for each request). In addition, three image examples were provided with each topic (these images were not removed from the collection, but removed from the set of relevance judgments). The topic titles were then translated into 15 languages including German, French, Spanish, Italian, Portuguese, Dutch, Russian, Japanese, and Simplified and Traditional Chinese. All translations were provided by at least one native speaker and verified by at least another native speaker. Unlike in past campaigns, however, the topic narratives were neither translated nor evaluated this year.

1.3 Relevance Assessments

Relevance assessments were carried out by two topic creators⁷ using a custom-built online tool. The top 40 results from all submitted runs were used to create image pools giving an average of 1,045 images (max: 1468; min: 575) to judge per topic. The topic creators judged all images in the topic pools and also used interactive search and judge (ISJ) to supplement the pools with further relevant images (on average 25%). The ISJ was based on purely textual searches. The assessments were based on a ternary classification scheme: (1) relevant, (2) partially relevant, and (3) not relevant. Based on these judgments, only those images judged relevant by both assessors were considered for the set of relevant images (qrels).

⁵ We asked experts in the field to rate these topics according to the following scheme: (1) CBIR will produce very bad or random results, (2) bad results, (3) average results, (4) good results and (5) very good results.

⁶ The FIRE system was used based on using all query images <http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html>

⁷ One of the topic generators a member of the vivventura travel company.

1.4 Participating Groups and Methods

A total of 36 groups registered for ImageCLEFphoto this year, with exactly one third of them submitting a total of 157 runs (all of which were evaluated). Of the 12 participating groups, four of them were new to ImageCLEF: Berkeley, CINDI, TUC and CELI. Table 1 summarises participating groups and the number of runs submitted by them. All groups (with the exception of RWTH) submitted a monolingual English run with the most popular languages appearing as Italian, Japanese and Simplified Chinese (see Table 2).

Table 1. Participating groups for ImageCLEFphoto

Group ID	Institution	Runs
Berkeley	University of California, Berkeley, USA	7
CEA-LIC2M	Fontenay aux Roses Cedex, France	5
CELI	CELI srl, Torino, Italy	9
CINDI	Concordia University, Montreal, Canada	3
DCU	Dublin City University, Dublin, Ireland	40
IPAL	IPAL, Singapore	9(+4)
NII	National Institute of Informatics, Tokyo, Japan	6
Miracle	Daedalus University, Madrid, Spain	30
NTU	National Taiwan University, Taipei, Taiwan	30
RWTH	RWTH Aachen University, Aachen, Germany	2(+2)
SINAI	University of Jaén, Jaén, Spain	12
TUC	Technische Universität Chemnitz, Germany	4

Overall 157 runs were submitted using a variety of approaches. Participants were asked to categorise their submissions according to the following dimensions: query language, annotation language (English or German), run type (automatic or manual), use of feedback or automatic query expansion, and modality (text only, image only or combined). Table 4 shows the overall results according to runs categorised by these dimensions. Most submissions made use of the image annotations (or metadata), with 8 groups submitting bilingual runs and 11 groups monolingual runs (many groups used MT systems for translation, e.g. Berkeley, DCU, NII, NTU and SINAI). For many participants, the main focus of their submission was combining visual and text features (11 groups submitted text-only runs; 7 used a combination of text and visual information: CEA, CINDI, DCU, IPAL, Miracle, NTU and TUC) and/or using some kind of relevance feedback to provide query expansion (8 groups using some kind of feedback: Berkeley, CINDI, DCU, IPAL, Miracle, NTU, SINAI and TUC).

Based on all submitted runs, 59% were bilingual (85% X-English; 15% X-German), 31% involved the use of image retrieval (27% using combined visual and textual features) and 46% of runs involved some kind of relevance feedback (typically in the form of query expansion). The use of query expansion was shown to increase retrieval effectiveness by bridging the gap between the languages of the query and annotations. The majority of runs were automatic (i.e. involving no human intervention); 1 run was manual. Further details of methods used in the submitted runs can be found in the workshop papers submitted by participants.

Table 2. Ad-hoc experiments listed by query and annotation language

Query Language	Annotation	# Runs	# Participants
English	English	49	11
Italian	English	15	4
Japanese	English	10	4
Simplified Chinese	English	10	3
French	English	8	4
Russian	English	8	3
German	English	7	3
Spanish	English	7	3
Portuguese	English	7	3
Dutch	English	4	2
Traditional Chinese	English	4	1
Polish	English	3	1
Visual	English	1	1
German	German	8	4
English	German	6	3
French	German	3	1
Japanese	German	1	1
Visual	(none)	6	3
Visual Topics	(none)	6	2

1.5 Results and Discussion

Analysis of System Runs. Results for submitted runs were computed using version 7.3 of `trec_eval`⁸. Submissions were evaluated using uninterpolated (arithmetic) Mean Average Precision (MAP) and Precision at rank 20 (P20). We also considered Geometric Mean Average Precision (GMAP) to test robustness [6]. Using Kendall's Tau to compare system ranking between measures, significant correlations existed at $p \leq 0.01$ between all measures above 0.74. This implies that the measure used to rank systems *does* affect system ranking and requires further investigation.

Table 3 shows the runs which achieved the highest MAP for each language pair. Of these runs, 83% use feedback of some kind (typically pseudo relevance feedback) and a similar proportion use both visual and textual features for retrieval. It is interesting to note that English monolingual outperforms the German monolingual (19% lower) and the highest bilingual to English run was Portuguese-English which performed 74% of monolingual, but the highest bilingual to German run was English to German which performed only at only 39% of monolingual. Also, unlike previous years, the top-performing bilingual runs involved Portuguese, traditional Chinese and Russian as the source language showing an improvement of the retrieval methods using these languages.

Table 4 shows results by different dimensions and indicates that, on average, combining visual features from the image and semantic information from the annotations gives a 54% improvement over text alone, using some kind of feedback (visual and textual) gives a 39% improvement, bilingual retrieval performs 7% lower than monolingual and that results for English as the target language are 26% higher than those for German (differences are statistically significant at $p \leq 0.05$ using a Student's t-Test). The differences are less impressive, however,

⁸ http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz

Table 3. Systems with highest MAP for each query language (ranked by descending order of MAP scores)

Language (Captions)	Group	Run ID	MAP	P20	GMAP
English (English)	CINDI	Cindi_Exp_RF	0.385	0.530	0.282
German (German)	NTU	DE-DE-AUTO-FB-TXTIMG-T-WEprf	0.311	0.335	0.132
Portuguese (English)	NTU	PT-EN-AUTO-FB-TXTIMG-T-WEprf	0.285	0.403	0.177
T. Chinese (English)	NTU	ZHS-EN-AUTO-FB-TXTIMG-T-Ont-WEprf	0.279	0.464	0.154
Russian (English)	NTU	RU-EN-AUTO-FB-TXTIMG-T-WEprf	0.279	0.408	0.153
Spanish (English)	NTU	SP-EN-AUTO-FB-TXTIMG-T-WEprf	0.278	0.407	0.175
French (English)	NTU	FR-EN-AUTO-FB-TXTIMG-T-WEprf	0.276	0.416	0.158
Visual (English)	NTU	AUTO-FB-TXTIMG-WEprf	0.276	0.448	0.107
S. Chinese (English)	NTU	ZHS-EN-AUTO-FB-TXTIMG-T-WEprf	0.272	0.392	0.168
Japanese (English)	NTU	JA-EN-AUTO-FB-TXTIMG-T-WEprf	0.271	0.402	0.170
Italian (English)	NTU	IT-EN-AUTO-FB-TXTIMG-T-WEprf	0.262	0.398	0.143
German (English)	DCU	combTextVisual_DEENEN	0.189	0.258	0.070
Dutch (English)	DCU	combTextVisual_NLENEN	0.184	0.234	0.063
English (German)	DCU	combTextVisual_ENDEEN	0.122	0.175	0.036
Polish (English)	Miracle	miratctdplen	0.108	0.139	0.005
French (German)	DCU	combTextVisual_FRDEEN	0.104	0.147	0.002
Visual (none)	RWTHi6	RWTHi6-IFHTAM	0.063	0.182	0.022
Japanese (German)	NII	mcp.bl_jpn_ldger_ld.skl_dir	0.032	0.051	0.001

Table 4. MAP scores for each result dimension

Dimension	Type	# Runs	# Groups	Mean (σ)	Median	Highest
Query Language	bilingual	93	8	0.144 (0.074)	0.143	0.285
	monolingual	57	11	0.154 (0.090)	0.145	0.385
	visual	7	3	0.074 (0.090)	0.047	0.276
Annotation	English	133	11	0.152 (0.082)	0.151	0.385
	German	18	4	0.121 (0.070)	0.114	0.311
	none	6	2	0.041 (0.016)	0.042	0.063
Modality	Text Only	108	11	0.129 (0.062)	0.136	0.375
	Text + Image	43	7	0.199 (0.077)	0.186	0.385
	Image Only	6	2	0.041 (0.016)	0.042	0.063
Feedback/Expansion	without	85	11	0.128 (0.055)	0.136	0.334
	with	72	8	0.165 (0.090)	0.171	0.385

if the *highest* MAP score is considered: the highest bilingual run performs 35% lower than the monolingual, combining text and image gives only 3% increase and using feedback of some kind gives a 15% increase.

Absolute retrieval results are lower than previous years and we attribute this to the choice of topics, a more visually challenging photographic collection and there being incomplete annotations provided with the collection. All groups have shown that combining visual features from the image and semantic knowledge derived from the captions offers optimum retrieval for many of the topics. In general, feedback (typically in the form of query expansion based on pseudo relevance feedback) also appears to work well on short captions (including results from previous years) and is likely due to the limited vocabulary exhibited by the captions.

Analysis of Topics. There are considerable differences between the retrieval effectiveness of individual topics. Possible causes for these differences include: the discriminating power of query terms in the collection, the complexity of topics

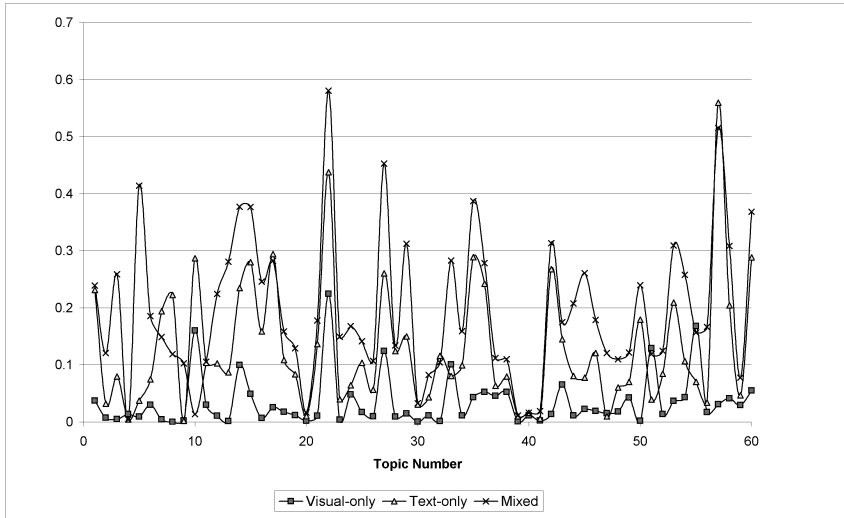


Fig. 2. MAP by topic based on modality

(e.g. topic 9 “Tourist accommodation near Lake Titicaca” involves a location and fuzzy spatial operator which would not be handled appropriately unless support for spatial queries was provided), the level of semantic knowledge required to retrieve relevant images (this limits the success of purely visual approaches), and translation success (e.g. whether proper names had been successfully handled). Fig. 2 shows mean average precision across (all) system runs for each topic based on modality. Many topics clearly show improvement through the use of combining textual and visual features (mixed) than any single modality alone. Part of this is likely to be attributed to the availability of visual examples with the topics which could be used in the mixed runs (and that these examples were also present in the collection).

We are still investigating the effects of various retrieval strategies (e.g. use of visual and textual features and relevance feedback) on the results for different topics. We expect that the use of visual techniques will improve topics which can be considered “more visual” (e.g. “sunset over water” is more visual than “pictures of female guides” which one could consider more semantic) and that topics which are considered “more difficult” linguistically (e.g. “bird flying” is linguistically simpler than “pictures taken on Ayers Rock”) will require more complex language processing techniques.

1.6 The ImageCLEFphoto Visual Retrieval Sub-task

To investigate further the success of visual techniques, thirty topics from the ImageCLEFphoto task were selected and modified to reduce semantic information and make better suited to visual retrieval techniques. For example removing

Table 5. The visual results

RK	RUN ID	MAP	P20	GMAP
1	RWTHi6-IFHTAM	0.1010	0.2850	0.0453
2	RWTHi6-PatchHisto	0.0706	0.2217	0.0317
3	IPAL-LSA3-VisualTopics	0.0596	0.1717	0.0281
4	IPAL-LSA2-VisualTopics	0.0501	0.1800	0.0218
5	IPAL-LSA1-VisualTopics	0.0501	0.1650	0.0236
6	IPAL-MF-VisualTopics	0.0291	0.1417	0.0119

geographic constraints (e.g. “black and white photos” instead of “black and white photos *from Russia*”) and other, non-visual constraints (e.g. “*child* wearing baseball cap” instead of “*godson* wearing baseball cap”). We wanted to attract more visually orientated groups to ImageCLEFphoto which to date has been dominated by groups using textual approaches.

The same document collection was used as with the ImageCLEFphoto task, but without the corresponding image captions. Participants were given three example images to describe each topic and were required to perform query-by-visual-example retrieval to begin the search. Two out of 12 groups that participated at the general ImageCLEFphoto task also submitted a total of six runs for the visual subtask: IPAL and RWTH.

Relevance judgments were performed as described in Section 1.3: the top 40 results from the six submitted runs were used to create image pools giving an average of 171 images (max: 190; min: 83) to judge per topic. The topic creators judged all images in the topic pools and used (text-based) ISJ to supplement the pools with further relevant images.

Most runs had quite promising results for precision values at a low cut-off ($P20 = 0.285$ for the best run, compare the results shown in Table 5). However, it is felt that this is due to the fact that some relevant images in the database are visually very similar to the query images, rather than algorithms really understanding what is being searched for. The retrieved images at higher ranks appeared random and further relevant images were only found by chance. This is also reflected by the low MAP scores (0.101 for the best run). Image retrieval systems typically achieve good results for tasks based on specific domains, or in tasks which are well-suited to the current level of CBIR. The low results of the visual sub-task highlight the fact that the successful application of visual techniques in applications which involve more general (and less domain-specific) pictures is still requiring much investigation.

It has to be further investigated with the participants why only two (out of 36 registered) groups actually submitted visual-only results. On the one hand, some groups mentioned in their feedback that they could not submit due to lack of time; the generally low results for this task might have also discouraged several groups from submitting their results. On the other hand, there were twice as many groups that submitted purely content-based runs to the main ImageCLEFphoto task; the question might arise whether this visual task has been promoted sufficiently enough and it should further be discussed with participants.

2 The Object Annotation Task

After the success of the automatic medical annotation task in 2005 [7] that clearly showed the need for public evaluation challenges in computer vision, and several calls for a more general annotation task from ImageCLEF participants, a plan for a non-medical automatic image classification (or annotation) task was created. In contrast to the medical task, images to be labeled were of everyday objects and hence did not require specialised domain knowledge. The aim of this new annotation task was to identify objects shown in test images and label the image accordingly. In contrast to the PASCAL visual object classes challenge⁹ [8] where several two-class experiments are performed, i.e. independent prediction of presence or absence of various object classes, in this task several object classes were tackled jointly.

2.1 Database and Task Description

LTUtech¹⁰ kindly provided their hand collected dataset of images from 268 classes. Each image of this dataset shows one object in a rather clean environment, i.e. the images show the object and some, mostly homogeneous, background.

To facilitate participation in the first year, the number of classes were reduced to 21. The classes 1) “*ashtrays*”, 2) “*backpacks*”, 3) “*balls*”, 4) “*banknotes*”, 5) “*benches*”, 6) “*books*”, 7) “*bottles*”, 8) “*cans*”, 9) “*calculators*”, 10) “*chairs*”, 11) “*clocks*”, 12) “*coins*”, 13) “*computer equipment*”, 14) “*cups and mugs*”, 15) “*hifi equipment*”, 16) “*cutlery(knives, forks and spoons)*”, 17) “*plates*”, 18) “*sofas*”, 19) “*tables*”, 20) “*mobile phones*”, and 21) “*wallets*” were used. Removing all images that did not belong to one of these classes lead to a database of 13,963 images. To create a new set of test data, 1,100 new images of objects from these classes were taken. In these images, the objects were in a more “natural setting”, i.e. there was more background clutter than in the training images. To simplify the classification task, it was specified in advance that each test image belonged to only one of the 21 classes. Multiple objects of the same class may appear in an image. Objects not belonging to any of the 21 classes may appear as background clutter.

The training data was released together with 100 randomly sampled test images with known classifications to allow for the tuning of system parameters. Following this, the remaining 1000 images were published unclassified as the test data.

The distribution of classes in both the training and test data was non-uniform. Table 6 summaries the distributions and Figure 3 shows examples from the training and test data for each of the classes. It can be seen from the images that the task is hard. This is because the test data contains far more clutter than the training data.

⁹ <http://www.pascal-network.org/challenges/VOC/>

¹⁰ <http://www.ltutech.com>

Table 6. Overview of the data of the object annotation task

class	train	dev	test
1 Ashtrays	300	1	24
2 Backpacks	300	3	28
3 Balls	320	3	10
4 Banknotes	306	4	45
5 Bench	300	1	44
6 Books	604	5	65
7 Bottles	306	9	95
8 Calculators	301	1	14
9 Cans	300	0	20
10 Chairs	320	10	132
11 Clocks	1833	2	47
12 Coins	310	0	26
13 Computing equipment	3923	10	79
14 Cups	600	12	108
15 HiFi	1506	2	24
16 Cutlery	912	12	86
17 Mobile Phones	300	6	39
18 Plates	302	9	52
19 Sofas	310	3	22
20 Tables	310	2	23
21 Wallets	300	5	17
sum	13963	100	1000

2.2 Participating Groups and Methods

20 groups registered for the general annotation task and 3 of these submitted a total of 8 runs. Details about each of the participating groups and their submissions is provided in the following text (groups are listed alphabetically by their group id, which is later used in the results section to refer to the groups):

- *CINDI*. The CINDI group from Concordia University in Montreal, Canada, submitted 4 runs. For their experiments they used MPEG7 edge direction histograms and MPEG7 color layout descriptors, classified by a nearest neighbour classifier and by different combinations of support vector machines. They expected their run **SVM-Product** to be their best submission.
- *DEU*. The DEU group from the Department of Computer Engineering of the Dokuz Eylul University in Tinaztepe, Turkey, submitted 2 runs. For their experiments they used MPEG7 edge direction histograms and MPEG7 colour layout descriptors respectively. For classification, a nearest prototype approach was taken.
- *RWTHi6*. The Human Language Technology and Pattern Recognition Group from the RWTH Aachen University in Aachen, Germany, submitted 2 runs. For image representation they used a bag-of-features approach and for classification a discriminatively trained maximum entropy (log-linear) model was used. Their runs differed with respect to the histogram bins and vector quantization methods chosen.
- *MedGIFT*. The MedGIFT group of the University and Hospitals of Geneva, Switzerland, submitted two runs. One run was entirely based on tf/idf weighting of the GNU Image Finding Tool (GIFT). This acted as a baseline using only collection frequencies of features with no learning on the training data

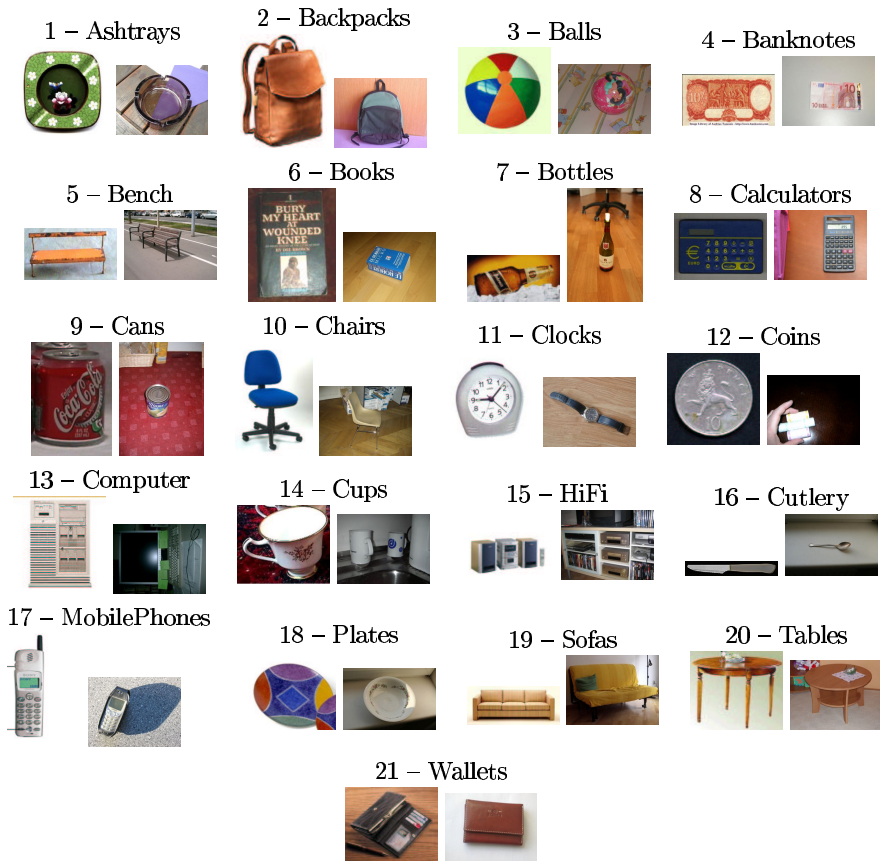


Fig. 3. One image from training (left) and test (right) data for each of the classes

supplied. The other submission was a combination of several separate runs by voting. The single results were quite different, so the combination-run was expected to be the best submission. The runs were submitted after the evaluation ended and therefore not evaluated (or ranked).

2.3 Results

The results of the evaluation are shown in Table 7. The runs are sorted by error rate. Overall, the error rates are very high due to the difficulty of the task. Scores range from 77.3% to 93.2%, indicating that many of the test images could not be classified correctly using any method. Table 7 provides details on the number of correctly classified images. None of the test images were classified correctly by all classifiers; 411 images were misclassified by all submitted runs and 301 images could be classified correctly by only one classifier.

Table 7. Results from the object annotation task sorted by error rate

rank	Group ID	Runtag	Error rate
1	RWTHi6	SHME	77.3
2	RWTHi6	PatchHisto	80.2
3	cindi	Cindi-SVM-Product	83.2
4	cindi	Cindi-SVM-EHD	85.0
5	cindi	Cindi-SVM-SUM	85.2
6	cindi	Cindi-Fusion-knn	87.1
7	DEU-CS	edgehistogr-centroid	88.2
-	medGIFT	fw-bwpruned	90.5
-	medGIFT	baseline	91.7
8	DEU-CS	colorlayout-centroid	93.2

Table 8. The number of test images correctly classified by the number of runs

number of images	number of runs in which correctly classified
411	0
301	1
120	2
69	3
54	4
30	5
13	6
2	7
0	8

We have found that a combination of classifiers can improve results. Using the first two methods and summing up normalized confidences leads to an error rate of 76.7%, and using the three best submissions leads to an error rate of 75.8%. Adding further submissions could not improve the performance further. Combining all submissions lead to an error rate of 78.8%.

2.4 Discussion

Considering that the error rates of the submitted runs are high and that nearly half of these images could not be classified correctly by any of the submitted methods, it can be said that the task was very challenging. One contributing factor was that the training images generally contained very little clutter and obscuring features, whereas the test images showed objects in their “natural” (and more realistic) environment. None of the participating groups specifically addressed this issue, although it would be expected to lead to improvement in classification accuracy. Furthermore, the results show that discriminatively trained methods outperform other methods (similar to results from the medical automatic annotation task), although only by a small amount that is probably not statistically significant.

Although the object annotation task and the medical automatic annotation tasks of ImageCLEF 2006 [9] were similar, they also differed in four important ways:

- Both tasks provided a relatively large training set and a disjunct test set. Thus, in both cases it is possible to learn a relatively reliable model for the training data (this is somewhat proven for the medical annotation task).
- Both tasks were multi-class/one object per image classification tasks. Here they differ from the PASCAL visual classes challenge which has addressed a set of object vs. non object tasks where several objects (of equal or unequal type) may be contained in an image.
- The medical annotation task has only gray scale images, whereas the object task has mainly color images. This is probably most relevant for the selection of descriptors.
- The images from the test and the training set were from the same distribution for the medical task, whereas for the object task the training images are rather clutter-free and the test images contained a significant amount of clutter. This is probably relevant and should be addressed when developing methods for the non-medical task. Unfortunately, the participating methods did not address this issue which probably has a significant impact on the results.

3 Conclusions

ImageCLEF continues to provide resources to the retrieval and computational vision communities to facilitate standardised laboratory-style testing of (pre-dominately text-based) image retrieval systems. The main division of effort thus far in ImageCLEF has been between medical and non-medical information systems. These fields have helped to attract different groups to ImageCLEF (and CLEF) over the past 2-3 years and thereby broaden the audience of this evaluation campaign. For the retrieval task, the first 2 evaluation events were based on cross-language retrieval from a cultural heritage collection: the St Andrews historic collection of photographic images. This provided certain challenges for both the text and visual retrieval communities, most noticeably the style of language used in the captions and the types of pictures in the collection: mainly black-and-white of varying levels of quality and visual degradation.

For 2006, the retrieval task moved to a new collection based on feedback from ImageCLEF participants in 2005-2006 and the availability of the IAPR-TC12 Benchmark¹¹. Designed specifically as a benchmark collection, it is well-suited for use in ImageCLEF with captions in multiple languages and high-quality colour photographs covering a range of topics. This type of collection - personal photographs - is likely to become of increasing interest to researchers with the growth of the desktop search market and popularity of tools such as Flickr¹².

Like in previous years, the ImageCLEFphoto task has shown the usefulness of combining visual and textual features derived from the images themselves

¹¹ One of the biggest factors influencing what collections are used and provided by ImageCLEF is copyright.

¹² <http://www.flickr.com>

and associated image captions (although to a lesser degree this year). It is noticeable that, although some topics are more “visual” than others and likely to benefit more from visual techniques, the majority of topics seem to benefit from a combination of text and visual approaches and participants continue to deal with issues involved in combining this evidence. In addition, the use of relevance feedback to facilitate, for example, query expansion in text retrieval continues to improve the results of many topics in collections used so far, likely due to the nature of the text associated with images: typically a controlled vocabulary that lends itself to blind relevance feedback.

For the automatic annotation/object classification task the addition of the LTU dataset has provided a more general challenge to researchers than medical images. The object annotation task has shown that current approaches to image classification and/or annotation have problems with test data that is not from the same distribution as the provided training data. Given the current high interest in object recognition and annotation in the computer vision community it is to be expected that big improvements are achievable in the area of automatic image annotation in the near future. It is planned to use image annotation techniques as a preprocessing step for a multi-modal information retrieval system: given an image, create an annotation and use the image and the generated annotation to query a multi-modal information retrieval system, which is likely to improve the results given the much better performance of combined runs in the photographic retrieval task.

Acknowledgements

We would like to thank *viventura*, the IAPR and LTUtech for providing their image databases for this years’ tasks, and to Tobias Weyand for creating the web interface for submissions. This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contracts NE-572/6 and Le-1108/4, the Swiss National Science Foundation (FNS) under contract 205321-109304/1, the American National Science Foundation (NSF) with grant ITR-0325160, an International Postgraduate Research Scholarship (IPRS) by Victoria University, the EU SemanticMining project (IST NoE 507505) and the EU MUSCLE NoE.

References

1. Voorhees, E.M., Harman, D.: Overview of the seventh Text REtrieval Conference (TREC-7). In: *The Seventh Text Retrieval Conference*, Gaithersburg, MD, USA, pp. 1–23 (1998)
2. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross-language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004*. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)

3. Grubinger, M., Clough, P., Müller, H., Deseleers, T.: The IAPR-TC12 benchmark: A new evaluation resource for visual information systems. In: International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, held in conjunction with LREC'06, Genoa, Italy, pp. 13–23 (2006)
4. Zhang, V., Rey, B., Stipp, E., Jones, R.: Geomodification in query rewriting. In: GIR '06: Proceedings of the Workshop on Geographic Information Retrieval, SIGIR (2006)
5. Grubinger, M., Leung, C., Clough, P.D.: Linguistic estimation of topic difficulty in cross-language image retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 558–566. Springer, Heidelberg (2006)
6. Voorhees, E.M.: The trec robust retrieval track. SIGIR Forum 39, 11–20 (2005)
7. Müller, H., Geissbühler, A., Marty, J., Lovis, C., Ruch, P.: The Use of medGIFT and easyIR for ImageCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J.D.R., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J.: The 2005 pascal visual object classes challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 117–176. Springer, Heidelberg (2006)
9. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In: CLEF working notes, Alicante, Spain (2006)

Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks

Henning Müller¹, Thomas Deselaers², Thomas Deserno³, Paul Clough⁴,
Eugene Kim⁵, and William Hersh⁵

¹ Medical Informatics, University and Hospitals of Geneva, Switzerland

² Computer Science Dep., RWTH Aachen University, Germany

³ Medical Informatics, RWTH Aachen University, Germany

⁴ Sheffield University, Sheffield, UK

⁵ Oregon Health and Science University (OHSU), Portland, OR, USA
`henning.mueller@sim.hcuge.ch`

Abstract. This paper describes the medical image retrieval and annotation tasks of ImageCLEF 2006. Both tasks are described with respect to goals, databases, topics, results, and techniques. The ImageCLEFmed retrieval task had 12 participating groups (100 runs). Most runs were automatic, with only a few manual or interactive. Purely textual runs were in the majority compared to purely visual runs but most were mixed, using visual and textual information. None of the manual or interactive techniques were significantly better than automatic runs. The best-performing systems used visual and textual techniques combined, but combinations of visual and textual features often did not improve performance. Purely visual systems only performed well on visual topics. The medical automatic annotation used a larger database of 10,000 training images from 116 classes, up from 9,000 images from 57 classes in 2005. Twelve groups submitted 28 runs. Despite the larger number of classes, results were almost as good as in 2005 which demonstrates a clear improvement in performance. The best system of 2005 would have received a position in the middle in 2006.

Keywords: image retrieval, automatic image annotation, medical information retrieval.

1 Introduction

ImageCLEF¹ [1] started within CLEF (Cross Language Evaluation Forum) in 2003. A medical image retrieval task was added in 2004 to explore domain-specific retrieval as well as multi-modal retrieval (combining visual and textual features for retrieval). Since 2005, a medical retrieval and a medical image annotation task have been parts of ImageCLEF. This paper concentrates on the two medical tasks, whereas a second paper [2] describes the new object classification and the photographic retrieval tasks. More detailed information can also be

¹ <http://ir.shef.ac.uk/imageclef/>

found on the task web pages for ImageCLEFmed² and the medical annotation task³. Detailed analyses of the 2005 medical image retrieval task and of the 2005 medical annotation task are available in [3] and [4], respectively.

2 The Medical Image Retrieval Task

2.1 General Overview

In 2006, the medical retrieval task was run for the third year, and for the second year with the same dataset of over 50,000 images from four collections. One of the most interesting findings for 2005 was the variable performance of systems based on whether the topics had been classified as amenable to visual, textual, or mixed retrieval methods. For this reason, we developed 30 topics for 2006, with 10 each in the three categories. The scope of the topic development was slightly enlarged by using the log files of a medical media search engine of the Health on the Net (HON) foundation. Analysis of these logs showed a great number of general topics not covering the entire four axes defined in 2005:

- Anatomic region shown in the image;
- Image modality (e.g. x-ray, CT, MRI, gross pathology, etc.);
- Pathology or disease shown in the image;
- Abnormal visual observation (e.g. enlarged heart).

The process of relevance judgements was similar to 2005 and `trec_eval` was used for the evaluation of the results.

2.2 Registration and Participation

In 2006, a record number of 47 groups registered for ImageCLEF and among these, 37 registered for the medical retrieval task. Groups came from four continents and a total of 16 countries. Unfortunately, some registered group did not send in results. 12 groups from 8 countries submitted results. Each entry below describes briefly the techniques used for their submissions.

- *CINDI, Canada*. The CINDI group from Concordia University, Canada, submitted a total of four runs: one purely textual, one visual, and two combined runs. Text retrieval was based on Apache Lucene. For visual information a combination of global and local features were used and compared using Euclidean distance. Most submissions used relevance feedback (RF).
- *MSR, China*. Microsoft Research China submitted one purely visual run using a combination of various features accounting for color and texture.

² <http://ir.ohsu.edu/image/>

³ <http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef06/mediclaat.html>

- *Institute for Infocomm Research I2R-IPAL, Singapore.* IPAL submitted 26 runs, the largest number of any group. Textual and visual runs were prepared in cooperation with I2R. For visual retrieval, patches of image regions were applied and manually classified into semantically valid categories and mapped to Unified Medical Language System (UMLS). For the textual analysis, all query languages were separately mapped to UMLS and then applied to retrieval. Several classifiers based on SVMs and other classical approaches were used and combined.
- *UKLFR, Germany.* The University Hospitals of Freiburg submitted 9 runs mainly using textual retrieval. Interlingua and the original language were used (morphosaurus and Lucene). Queries were preprocessed by removing the “show me” test. Runs differed in language and combination with GIFT.
- *SINAI, Spain.* Jaen University submitted 12 runs: 3 of them using only textual information and 9 using a text retrieval system and adding provided data from the GIFT image retrieval system. The runs differed in settings for “information gain” and the weighting of textual and visual information.
- *OHSU, USA.* Oregon Health and Science University performed manual modification of queries and fusion with results from visual runs. One run established a baseline using the text of the topics as given. Another run then manually modified the topic text removing common words and adding synonyms. For both runs, there were submissions in each of the three individual languages (English, French, German) plus a merged run with all and a run with the English topics expanded with automatic translation using the Babelfish translator. The manual modification of the queries improved performance substantially. The best results came from the English-only queries, followed by the automatically translated and the merged queries. One additional run assessed fusing data from a visual run with the merged queries. This decreased MAP but did improve precision at 10 and 30 images.
- *I2R Medical Analysis Lab, Singapore.* Their submission was together with the IPAL group from the same lab.
- *MedGIFT, Switzerland.* The University and Hospitals of Geneva relied on two retrieval systems for their submission. The visual part was performed with medGIFT. The textual retrieval used a mapping of the query and document text towards concepts in MeSH (Medical Subject Headings). Then, matching was performed with a frequency-based weighting method using easyIR. All results were automatic runs using visual, textual and mixed features. Separate runs were submitted for the three languages.
- *RWTH Aachen University – Computer Science, Germany.* The RWTH Aachen University, CS department, submitted 9 runs, all using the FIRE system and various features describing color, texture, and global appearance. For one run, the queries and the qrels of last year were used as training data to obtain weights for the combination of features using maximum entropy training. One run was purely textual, 3 were purely visual, and the remaining 5 runs used textual and visual information. All runs were fully-automatic.
- *RWTHmi, Germany.* The medical Informatics group at RWTH Aachen submitted 2 purely visual runs without interaction. Both runs used a combination

of global appearance and texture features compared with invariant distance measures. Runs differed in the weights for the features.

- *SUNY, USA*. State University New York submitted 2 purely textual runs and 2 using text and visual information. Parameters for the system were tuned using 2005 topics and automatic RF in variations.
- *LITIS Lab, INSA Rouen, France*. The INSA group from Rouen submitted one run using visual and textual information. MeSH dictionaries were used for text analysis, and the images were represented by various features accounting for global and local information. Most of the topics were treated fully automatic and four topics were treated with manual interaction.

2.3 Databases

In 2006, the same dataset was used as in 2005 containing four sets of images. Casimage was made available, containing almost 9,000 images of 2,000 cases [5]. Casimage includes mostly radiology, but also photographs, PowerPoint slides, and illustrations. Cases are mainly in French, with around 20% being in English and 5% without annotation. We also used PEIR⁴ (Pathology Education Instructional Resource) with annotations based on HEAL⁵ (Health Education Assets Library, mainly Pathology images [6]). This dataset contains over 33,000 images with English annotations, annotation being per image. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology [7], was also distributed containing over 2,000 images mainly from nuclear medicine with annotations provided per case and in English. Finally, PathoPic [8] was included in our dataset containing 9,000 images with extensive annotation per image in German. Part of the German annotation is translated into English. As such, we were able to use a total of more than 50,000 images, with annotations in three different languages. Through an agreement with the copyright holders, we were able to distribute these images to the participating research groups.

2.4 Query Topics

The query topics were based on two surveys performed in Portland and Geneva [9,10]. In addition to this, a log file of a media search engine HON⁶ was used to create topics along the following axes:

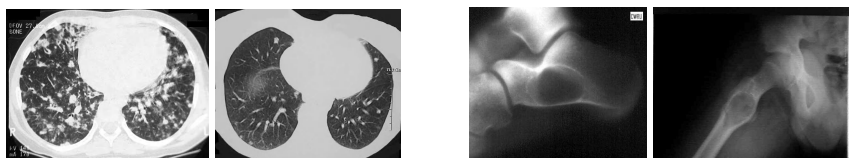
- Anatomic region shown in the image;
- Image modality (x-ray, CT, MRI, gross pathology, etc.);
- Pathology or disease shown in the image;
- Abnormal visual observation (e.g. enlarged heart).

The HON log-files indicated rather general topics than the specific ones used in 2005, so we used real queries from the log-files in 2006. We could not use the most frequent queries, since they were too general, e.g. heart, lung, etc.,

⁴ <http://peir.path.uab.edu/>

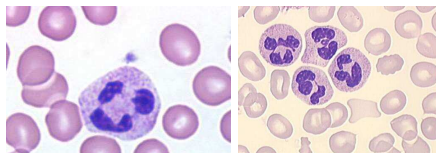
⁵ <http://www.healcentral.com/>

⁶ <http://www.hon.ch/>



Show me chest CT images with nodules.
Zeige mir CT Bilder der Lunge mit Knötchen.
Montre-moi des CTs du thorax avec nodules.

Show me x-ray images of bone cysts.
Zeige mir Röntgenbilder von Knochenzysten.
Montre-moi des radiographies de kystes d'os.



Show me blood smears that include polymorphonuclear neutrophils.
Zeige mir Blutabstriche mit polymorphonuklearer Neutrophils.
Montre-moi des échantillons de sang incluant des neutrophiles polymorphonucéaires.

Fig. 1. Examples for a visual, a mixed and a semantic topic

but those that satisfied at least two of the axes. After identifying 50 candidate topics, we grouped them into three classes based upon an estimation of what retrieval techniques they would be most retrievable – visual, mixed, or textual. Another goal was to cover frequent diseases and have a balanced variety of imaging modalities and anatomic regions. After choosing 10 queries for each category, we manually searched query images on the web. In 2005, images were taken partly from the collection. Although they were most often cropped, having external images made the visual task more challenging, as these images could be from other modalities and have completely different characteristics. Figure 1 shows examples for visual, mixed and semantic topics.

2.5 Relevance Judgements

For relevance judging, pools were built from all images for a given topic ranked in the top 30 retrieved. This gave pools from 647 to 1,187 images, with a mean of 910 per topic. Relevance judgements were performed by seven US physicians enrolled in the OHSU biomedical informatics graduate program. Eleven of the 30 topics were judged in duplicate, with two judged by three different judges. Each topic had a designated “original” judge from the seven. A total of 27,306 relevance judgements were made. (These were primary judgements; ten topics had duplicate judgements.) The judgements were turned into a qrels file, which was then used to calculate results with `trec_eval`. We used Mean Average Precision (MAP) as the primary evaluation measure. We note, however, that its orientation to recall (over precision) may may not be appropriate for many image retrieval tasks.

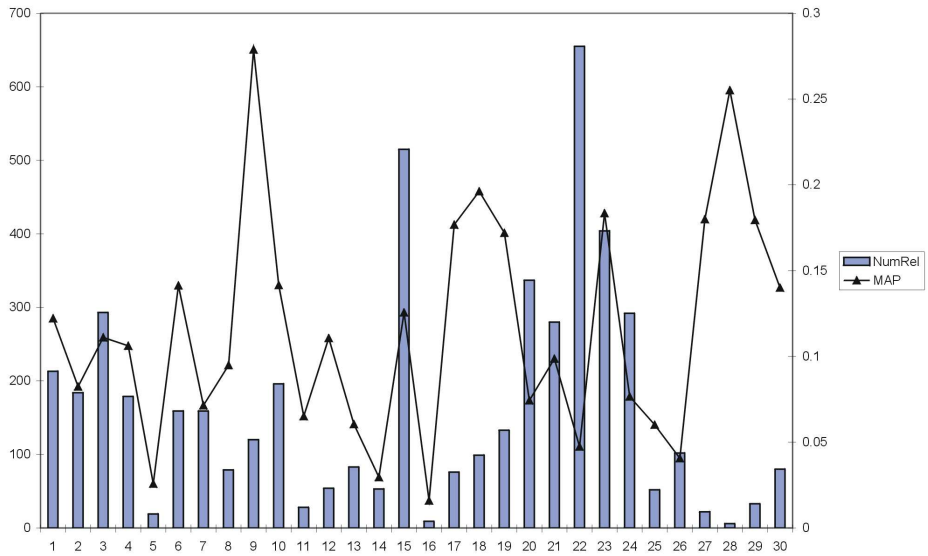


Fig. 2. Evaluation results and number of relevant images per topic

2.6 Submissions and Results

12 groups from eight countries participated in ImageCLEFmed 2006. These groups submitted 100 runs, with each group submitting from 1 to 26 runs.

We defined two categories for the submitted runs: one for the interaction used (automatic – no human intervention, manual – human modification of the query before the output of the system is seen, and interactive – human modification of the query after the output of the system is seen) and one for the data used for retrieval (visual, textual, or a mixture). The majority of submitted runs were automatic. There were fewer visual runs than there were textual and mixed runs.

Figure 2 gives an overview of the number of relevant images per topic and of the performance that this topic obtained on average (MAP). It can be seen that the variation in this case was substantial. Some topics had several hundred relevant images in the collection, whereas others only had very few. Likewise, performance could be extremely good for a few topics and extremely bad for others.

Figure 3 illustrates a comparison of several measurements for all submitted runs. When looking at early precision (P(30)) the variations were large, but slowly disappear for later precision (P(100)). All measures correlate fairly well.

Automatic Retrieval. The category of automatic runs was by far the most common category for result submissions. 79 out of the 100 submitted runs were in this category. In Table 1 the best run of each participating system per category is shown.

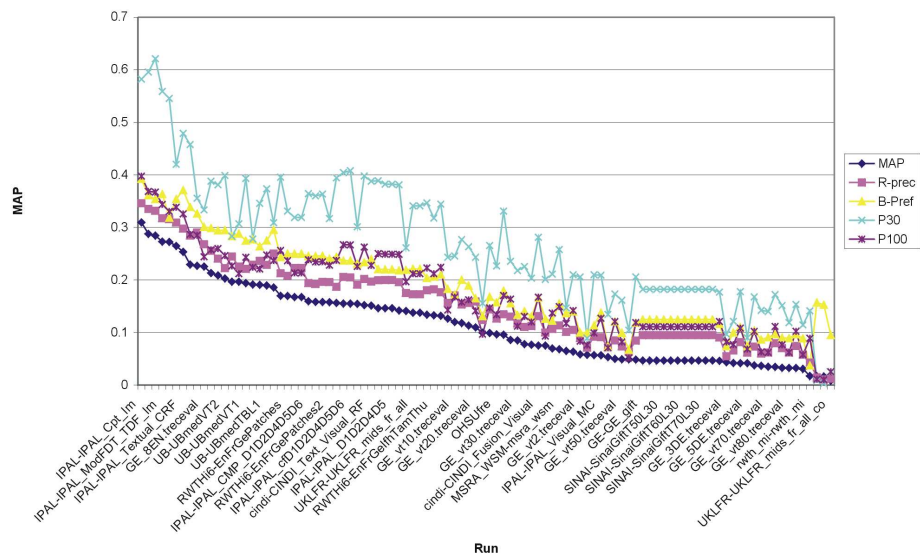


Fig. 3. Results for best runs of each system in each category, ordered by MAP

We can see that the best submitted automatic run was a mixed run and that other mixed runs had very good results. Nonetheless, several of the very good results were textual only, so a generalisation does not seem completely possible. Visual systems had a fairly low overall performance, although for the ten visual topics, their performance was very good.

Manual and Interactive Retrieval. Figure 2 shows the submitted manual and interactive runs. With the small numbers of manual runs, generalisation is

Table 1. Overview of the automatic runs

Run identifier	visual	textual	MAP	R-Prec
IPAL_Cpt_Im	x	x	0.3095	0.3459
IPAL_Textual_CDW		x	0.2646	0.3093
GE_8EN.treceval		x	0.2255	0.2678
UB-UBmedVT2	x	x	0.2027	0.2225
UB-UBmedT1		x	0.1965	0.2256
UKLFR_origmids_en_en		x	0.1698	0.2127
RWTHi6-EnFrGePatches	x	x	0.1696	0.2078
RWTHi6-En		x	0.1543	0.1911
OHSU_baseline_trans		x	0.1264	0.1563
GE_vt10.treceval	x	x	0.12	0.1703
SINAI-SinaiOnlytL30		x	0.1178	0.1534
CINDI_Fusion_Visual	x		0.0753	0.1311
MSRA_WSM-msra_wsm	x		0.0681	0.1136
IPAL_Visual_SPC+MC	x		0.0634	0.1048
RWTHi6-SimpleUni	x		0.0499	0.0849
SINAI-SinaiGiftT50L20	x	x	0.0467	0.095
GE-GE_gift	x		0.0467	0.095
UKLFR_mids_en_all_co	x	x	0.0167	0.0145

Table 2. Overview of the manual and interactive runs

Run identifier manual	visual	textual	MAP	R-Prec
OHSUeng		x	0.2132	0.2554
IPAL_CMP_D1D2D4D5D6	x		0.1596	0.1939
INSA-CISMef	x	x	0.0531	0.0719
Run identifier interactive	visual	textual	MAP	R-Prec
IPAL_Textual_CRF		x	0.2534	0.2976
OHSU-OHSU_m1	x	x	0.1563	0.187
CINDI_Text_Visual_RF	x	x	0.1513	0.1969
CINDI_Visual_RF	x		0.0957	0.1347

difficult. The first interactive run had good performance but was still not better than the best automatic run of the same group.

2.7 Conclusions

The best overall run by the IPAL institute is an automatic run using visual and textual features. We can tell from the submitted runs that interactive and manual runs do not perform better than the automatic runs. This may be partly due to the fact that most groups submitted more automatic runs than other runs. The automatic approach appears to be less time-consuming and most research groups have more experience in optimising these runs. Visual features seem to be mainly good for the visual topics. Text-only runs perform very well, and only a few mixed runs manage to be better.

3 The Medical Automatic Annotation Task

Automatic image annotation is a classification task, where a given image is automatically labelled with a text describing its contents. In restricted domains, the annotation may be just a class from a constrained set of classes or it may be an arbitrary narrative text describing the contents of the images. In 2005, the medical automatic annotation task was performed in ImageCLEF to compare state-of-the-art approaches to automatic image annotation and classification [11]. This year’s medical automatic annotation task builds on top of last year: 1,000 new images were collected and the number of classes more than doubled, resulting in a harder task.

3.1 Database and Task Description

The complete database consists of 11,000 fully classified medical radiographs taken randomly from medical routine at the RWTH Aachen University Hospital. 9,000 of these were release together with their classification as training data, another 1,000 were also published with their classification as validation data to allow for tuning classifiers in a standardised manner. One thousand additional images were released at a later date without classification as test data. These 1,000 images had to be classified using the 10,000 images (9,000 training + 1,000 validation) as training data. The complete database of 11,000 images was

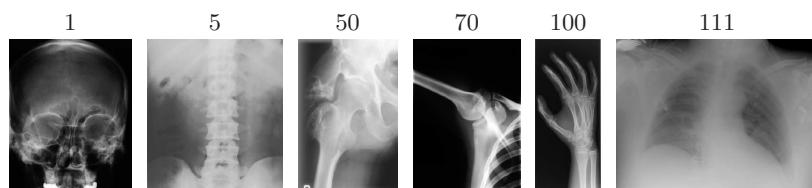


Fig. 4. Example images from the IRMA database with their class numbers

subdivided into 116 classes according to the IRMA code [12]. The IRMA code is a multi-axial code for the annotation of medical images. Currently, this code is available in English and German.

Example images from the database together with class numbers are given in Figure 4. The classes in the database are not uniformly distributed: class 111 has a 19.3% share of the complete dataset, class 108 has a 9.2% share, and 6 classes have only 1%/00 or less.

3.2 Participating Groups and Methods

27 groups registered and 12 of these submitted runs. Here, a short description of the methods of the submitted runs is provided. Groups are listed alphabetically by their group ID, which is later used in the results section to refer to the groups.

- *CINDI*. The CINDI group from Concordia University in Canada submitted 5 runs using a variety of features including MPEG-7 Edge Histogram Descriptor, MPEG-7 Color Layout Descriptor, invariant shape moments, downscaled images, and semi-global features. Some experiments combine these features with a PCA transformation. For 4 of the runs, a support vector machine (SVM) is used for classification with different multi-class voting schemes; in one run, the nearest neighbour decision rule is applied.
- *DEU*. The group of the Dokuz Eylul University in Tinaztepe, Turkey submitted one run which uses the MPEG-7 Edge Histogram as image descriptor and a 3-nearest neighbour classifier for classification.
- *MedIC-CISMeF*. The team from the INSA Rouen, France submitted 4 runs. Two use a combination of global and local image descriptors and the other two use local image descriptors. Features are dimensionality reduced by PCA and runs using the same features differing in the PCA coefficients are kept. The local features include statistical measures extracted from image regions and texture information. Yielding a 1953-dimensional feature vector when only local features are used and 2074-dimensional when local and global features are combined. For classification a SVM with RBF kernel is used.
- *MSRA*. The Web Search and Mining Group from Microsoft Research Asia submitted two runs. One run uses a combination of gray-block features, block-wavelet features, features accounting for binarised images, and an edge histogram. In total a 397-dimensional feature vector is used. The other run

uses a bag of features approach with vector quantisation where a histogram of quantised vectors is computed region-wise on the images. In both runs, SVMs are used for classification.

- *MU I2R*. The Media Understanding group of the Institute for Infocomm Research, Singapore submitted one run. A two-stage medical image annotation method was applied. First, the images are reduced to 32x32 pixels and classified using a SVM. Then, the decisions where the SVM was not sure, the decision was refined using a classifier that was trained on a subset of the training images. In addition to downscaled images, SIFT features and PCA transformed features were used for classification.
- *NCTU DBLAB*. The DBLAB of the National Chiao Tung University in Hsinchu, Taiwan submitted one run using tree image features, Gabor texture features, coherence moment and related vector layout as image descriptors. The classification was done using a nearest neighbour classifier.
- *OHSU*. The Department of Medical Informatics & Clinical Epidemiology of the Oregon Health and Science University in Portland, OR, USA submitted 4 runs. For image representation, a variety of descriptors was tested including 16x16 pixel versions of the images, and partly localised GLCM features. For classification multilayer perceptrons were used and settings were optimised using the development set.
- *RWTHi6*. The Human Language Technology and Pattern Recognition Group from the RWTH Aachen, Germany submitted 3 runs. One uses the image distortion model that was used for the best run of last year, and the other a sparse histogram of image patches and absolute position. The image distortion model run uses a nearest neighbour classifier, one of the other runs uses SVMs, and the other uses a maximum entropy classifier.
- *RWTHmi*. The IRMA group of the Medical Informatics Division of the RWTH Aachen University Hospital, Germany submitted two runs which using cross-correlation on 32x32 images with explicit translation shifts, image deformation model for Xx32 images, global texture features as proposed by Tamura, and global texture features as proposed by Castillo based on fractal concepts. For classification a nearest neighbour classifier was used. Weights for these features were optimised on the development set. One of these runs reflects their exact setup from 2005 for comparison.
- *UFR*. The Pattern Recognition and Image Processing group from the University Freiburg, Germany submitted two runs using gradient-like features extracted over interest points. Gradients over multiple directions and scale are calculated and used as a local feature vector. Features are clustered to form a codebook of size 20 and a cluster-cooccurrence matrix is computed over multiple distance ranges and multiple angle ranges (since rotation invariance is not desired), resulting in a 4D array per image which is flattened and used as the final feature vector. Classification is done using multi-class SVM in a one-vs-rest approach with a histogram intersection kernel.
- *ULG*. The group from the University of Liège, Belgium extracts a large number of possibly overlapping, square sub-windows of random sizes and at random positions from training images. Then, an ensemble model composed of

Table 3. Results of medical automatic annotation task; expected best marked with ‘*’

rank	Group	Runtag	Error rate [%]
* 1	RWTHi6	SHME	16.2
* 2	UFR	UFR-ns-1000-20x20x10	16.7
3	RWTHi6	SHSVM	16.7
4	MedIC-CISMeF	local+global-PCA335	17.2
5	MedIC-CISMeF	local-PCA333	17.2
6	MSRA	WSM-msra-wsm-gray	17.6
* 7	MedIC-CISMeF	local+global-PCA450	17.9
8	UFR	UFR-ns-800-20x20x10	17.9
9	MSRA	WSM-msra-wsm-patch	18.2
10	MedIC-CISMeF	local-PCA150	20.2
11	RWTHi6	IDM	20.4
* 12	RWTHmi	rwth-mi	21.5
13	RWTHmi	rwth-mi-2005	21.7
* 14	CINDI	cindi-svm-sum	24.1
15	CINDI	cindi-svm-product	24.8
16	CINDI	cindi-svm-ehd	25.5
17	CINDI	cindi-fusion-KNN9	25.6
18	CINDI	cindi-svm-max	26.1
* 19	OHSU	OHSU-iconGLCM2-tr	26.3
10	OHSU	OHSU-iconGLCM2-tr-de	26.4
21	NCTU	dblab-nctu-dblab2	26.7
22	MU	I2R-refine-SVM	28.0
23	OHSU	OHSU-iconHistGLCM2-t	28.1
* 24	ULG	SYSMOD-RANDOM-SUBWINDOWS-EX	29.0
25	DEU	DEU-3NN-EDGE	29.5
26	OHSU	OHSU-iconHist-tr-dev	30.8
27	UTD	UTD	31.7
28	ULG	SYSMOD-RANDOM-SUBWINDOWS-24	34.1

20 extremely randomised trees is automatically built based on size-normalised versions of the sub-windows, and operating directly on the pixel values to predict classes of sub-windows. Given this classifier a new image is classified by classifying sub-windows and combining classification decisions.

- *UTD*. The University of Texas, Dallas, USA submitted one run. Images are scaled to 16×16 pixels and their dimensionality is reduced by PCA. A weighted k-nearest neighbour algorithm is applied for classification.

3.3 Results

The results from the evaluation are given in Table 3. The error rates range from 16.2% to 34.1%. Based on the training data, a system guessing the most frequent group for all 1,000 test images would result with 80.5% error rate, since 195 radiographs of the test set were from class 111, which is the biggest class in the training data. A more realistic baseline is given by a nearest neighbour classifier using Euclidean distance to compare the images scaled to 32×32 pixels [13]. This classifier yields an error rate of 32.1%.

The average confusion matrix over all runs is given in Figure 5. It can clearly be seen that a diagonal structure is reached and thus that on the average most of the images were classified correctly. Some classes have high inter-class similarity: in particular classes 108 to 111 are often confused and in total many images from other classes were classified to be from class 111. Obviously, not all classes are

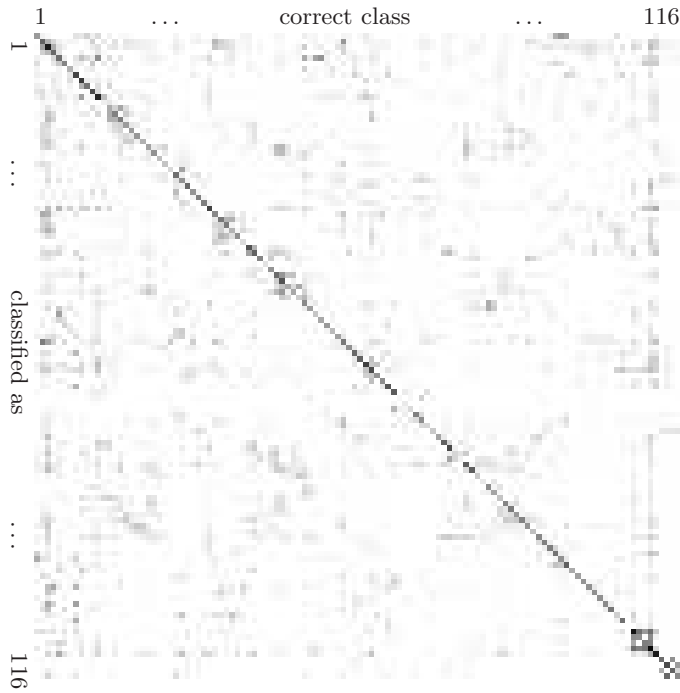


Fig. 5. Average confusion matrix over all runs. Dark points denote high entries, white points zero. X-axis shows the correct class and the Y-axis the class to which images were classified. Values are in logarithmic scale.

equally difficult, a tendency that classes with only few training instances are harder to classify than classes with a large amount of training data can be seen.

3.4 Discussion

The most interesting observation can be seen when comparing the results with those of last year: The RWTHi6-IDM [14] system that performed best in last years task (error rate: 12.1%) obtained an error rate of 20.4%. This increase in error rate can be explained by the larger number of classes and thus more similar classes that can easily be confused, on the other hand, 10 methods clearly outperform this result, 9 of these use SVMs as classifier (ranks 2-10) and one uses a discriminatively trained log-linear model (rank 1). Thus, it can clearly be said, that the performance of image annotation techniques strongly improved over one year, and that techniques that were initially developed for object recognition and detection are very well suited for the automatic annotation.

Given the confidence files of all runs, we tried to combine the classifiers by the sum rule. Therefore, all confidence files were normalised such that the confidences could be interpreted as a-posteriori probabilities $p(c|x)$ where c is the class and x

the observation. Unlike last year, where this technique could not improve results, clear improvements are possible combining several classifiers [15]: Using the top 3 ranked classifiers in combination, an error rate of 14.4% was obtained. The best result is obtained combining the top 7 ranked classifiers. No parameters were tuned but classifiers were combined equally.

4 Overall Conclusions

For the retrieval task, none of the manual or interactive techniques were significantly better than automatic runs. The best-performing systems used visual and textual techniques combined, but several times a combination of visual and textual features did not improve a system's performance. Thus, combinations for multi-modal retrieval need to be done carefully. Purely visual systems only performed well on the visual topics. For the automatic annotation task, discriminative methods outperformed methods based on nearest neighbour classification and the top-performing methods were based on the assumption that images consist of image parts, which can be modelled more or less independently.

One goal for future tasks is to motivate groups to work more on interactive and manual runs. Given enough manpower, such runs should be better than optimised automatic runs. Another future goal is to motivate an increasing number of subscribed groups to participate. Collections are planned to become larger. As some groups already complained about too large datasets for their computationally very expensive methods, a smaller database might be provided as an option for these groups to at least submit some results and compare them with other techniques. For the automatic annotation task, one goal is to use text labels with varying annotation precision rather than simple class-based annotation and to consider semi-automatic annotation methods.

Acknowledgements

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contracts NE-572/6 and Le-1108/4, the Swiss National Science Foundation (FNS) under contract 205321-109304/1, the American National Science Foundation (NSF) with grant ITR-0325160, and the EU 6th Framework Program with the SemanticMining (IST NoE 507505) and MUSCLE NoE.

References

1. Clough, P., Müller, H., Sanderson, M.: Overview of the CLEF cross-language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
2. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Proceedings of the CLEF 2006 workshop, Alicante, Spain. LNCS (September 2007)

3. Hersh, W., Müller, H., Jensen, J., Yang, J., Gorman, P., Ruch, P.: ImageCLEFmed: A text collection to advance biomedical image retrieval. *Journal of the American Medical Informatics Association*, 488–496 (2006)
4. Deselaers, T., Müller, H., Clogh, P., Ney, H., Lehmann, T.M.: The CLEF 2005 automatic medical image annotation task. *CLEF 2005* (2006)
5. Rosset, A., Müller, H., Martins, M., Dfouni, N., Vallée, J.-P., Ratib, O.: Casimage project – a digital teaching files authoring environment. *Journal of Thoracic Imaging* 19, 1–6 (2004)
6. Candler, C.S., Uijtdehaage, S.H., Dennis, S.E.: Introducing HEAL: The health education assets library. *Academic Medicine* 78, 249–253 (2003)
7. Wallis, J.W., Miller, M.M., Miller, T.R., Vreeland, T.H.: An internet-based nuclear medicine teaching file. *Journal of Nuclear Medicine* 36, 1520–1527 (1995)
8. Glatz-Krieger, K., Glatz, D., Gysel, M., Dittler, M., Mihatsch, M.J.: Web-basierte Lernwerkzeuge für die Pathologie – web-based learning tools for pathology. *Pathologie* 24, 394–399 (2003)
9. Hersh, W., Jensen, J., Müller, H., Gorman, P., Ruch, P.: A qualitative task analysis of biomedical image use and retrieval. In: *ImageCLEF/MUSCLE workshop on image retrieval evaluation*, Vienna, Austria, pp. 11–16 (2005)
10. Müller, H., Despont-Gros, C., Hersh, W., Jensen, J., Lovis, C., Geissbuhler, A.: Health care professionals' image use and search behaviour. In: *Proceedings of Medical Informatics Europe (MIE 2006)*, Maastricht, Netherlands, pp. 24–32 (2006)
11. Müller, H., Geissbuhler, A., Marty, J., Lovis, C., Ruch, P.: The Use of medGIFT and easyIR for Image CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 724–732. Springer, Heidelberg (2006)
12. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: *Proceedings SPIE*, vol. 5033, pp. 440–451 (2003)
13. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non-linear distortion models. In: *Proceedings BVM 2004, Bildverarbeitung für die Medizin*, Berlin, Germany, pp. 366–370 (2004)
14. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in Image-CLEF 2005: Combining content-based image retrieval with textual information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 688–698. Springer, Heidelberg (2006)
15. Kittler, J.: On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 226–239 (1998)

Text Retrieval and Blind Feedback for the ImageCLEFphoto Task

Ray R. Larson

School of Information
University of California, Berkeley, USA
`ray@sims.berkeley.edu`

Abstract. In this paper we will describe Berkeley’s approach to the ImageCLEFphoto task for CLEF 2006. This year is the first time that we have participated in ImageCLEF, and we chose to primarily establish a baseline for the Cheshire II system for this task, while we had originally hoped to use GeoCLEF methods for this task, in the end time constraints led us to restrict our submissions to the basic required runs for the task.

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley’s participation in the ImageCLEFphoto task [1]. Our submitted runs this year are intended to establish a simple baseline for comparison in future ImageCLEF tasks. This year we only used text-based retrieval methods for ImageCLEF, totally ignoring the images themselves (and the reference images specified in the queries). We hope, in future years, to be able to use combined text and image processing approaches similar to that used in earlier work combining Berkeley’s BlobWorld image retrieval system with the Cheshire II system (see [2]).

This year Berkeley submitted 7 runs, of which 4 were English monolingual, 2 German monolingual, and 1 bilingual English to German.

This paper first describes the retrieval methods used, including our blind feedback method for text, followed by a discussion of our official submissions and the results obtained from other methods after official submissions had been closed. Finally, we present some discussion of the results and our conclusions.

2 The Retrieval Algorithms

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions are discussed in detail in our GeoCLEF paper in this volume [3]. We used the “TREC2” logistic regression formula as well as blind feedback for our runs [3]. Since we lacked previous data on the behavior of our algorithms with the ImageCLEF data, we used the default values of the top-ranked 10 terms from the top-ranked 10 documents for blind feedback.

3 Approaches for ImageCLEF Indexing and Search Processing

Although the Cheshire II system uses the XML structure of documents and extracts selected portions of the record for indexing and retrieval, for the submitted runs this year we used only a single one of these indexes that contains the entire content of the document.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs, however, did *not* use decomposing in the indexing and querying processes to generate simple word forms from compounds (actually we tried, but there was a bug that failed to match any compounds in our runs).

Search processing for the ImageCLEF collection used Cheshire II scripts to parse the topics and submit the title or title and narrative from the topics to the “topic” index containing all terms from the documents. For the monolingual search tasks we used the topics in the appropriate language (English or German), and for bilingual tasks the topics were translated from the source language to the target language using SYSTRAN (via Babelfish at Altavista). We believe that other translation tools may provide a more accurate representation of the topics.

We tried two main approaches for searching, the first used only the topic text from the title element, the second included both titles and the narrative elements. In all cases the “topic” index mentioned above was used, and probabilistic searches were carried out. Two forms of the TREC2 logistic regression algorithm were used. One used the basic algorithm, and the other used the TREC2 algorithm with blind feedback using the top 10 terms from the 10 top-ranked documents in the initial retrieval.

Our official runs did not make use of the example images in queries, but after our official submissions we realized that this information could be used in a text-only system as well. This involved retrieving the associated metadata records from the example images included in the queries and using their titles and location information to expand the basic query. We tested this in unofficial “POST” runs that are described below following the discussion of the official results.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the officially submitted bilingual and monolingual runs for both English and German are shown in Table 1. Table 2 has a number of unofficial runs described in the next section. Table 3 compares some results from the submitted runs and the unofficial runs for the ImageCLEFphoto task.

Table 1. Submitted ImageCLEF Runs

Run Name	Description	Query	Feedback	MAP
BERK_BI_ENDE_TN_T2FB	Biling. English⇒German	Title+Narr	Y	0.1054
BERK_MO_DE_T_T2	Monoling. German	Title	N	0.1362
BERK_MO_DE_T_T2FB	Monoling. German	Title	Y	0.1422
BERK_MO_EN_T_T2	Monoling. English	Title	N	0.1720
BERK_MO_EN_T_T2FB	Monoling. English	Title	Y	0.1824
BERK_MO_EN_TN_T2	Monoling. English	Title,Narr	N	0.2356
BERK_MO_EN_TN_T2FB	Monoling. English	Title,Narr	Y	0.2392

Table 2. Unofficial POST ImageCLEF Runs (using image metadata query expansion)

Run Name	Description	Query	Feedback	MAP
POST_BI_DEEN_T	Biling. German⇒English	Title	Y	0.3114
POST_BI_ENDE_T	Biling. English⇒German	Title	Y	0.2738
POST_BI_ENDE_TN	Biling. English⇒German	Title+Narr	Y	0.2645
POST_BI_FREN_T	Biling. French⇒English	Title	Y	0.2971
POST_MO_DE_T	Monoling. German	Title	Y	0.2906
POST_MO_EN_T	Monoling. English	Title	Y	0.3147
POST_MO_EN_TN	Monoling. English	Title+Narr	Y	0.3562

Table 3. Comparison of Official and Unofficial POST ImageCLEF Runs

Description	Query	Official POST Percent		
		MAP	MAP	Improv.
Bilingual English⇒German	Title+Narr	0.1072	0.2645	146.74
monolingual German	Title	0.158	0.2906	83.92
monolingual English	Title	0.1824	0.3147	72.53
monolingual English	Title+Narr	0.2392	0.3562	48.91

5 Discussion and Conclusions

The officially submitted runs described above, when compared to participants using combined text and image methods, are not particularly distinguished. The German monolingual and bilingual runs suffer from the lack of decompounding in this version of the system. For the purpose of establishing a baseline performance for our system the runs are adequate. One interesting observation was that although the MAP for runs with blind feedback was higher than matching runs without it, the precision at low recall levels was lower than runs without feedback. We suspect that this is an effect of the small size of the metadata records available for this task when compared to the full-text in other collections. Use of the narrative for English topics allowed us to further improve performance in the official runs, although the same pattern for feedback versus no feedback

performance as mentioned above was observed, indicative that the pattern is likely a function of the record size and not query size.

Although our official runs for ImageCLEFphoto 2006 are in no way outstanding, they do provide a “foot in the door” and a baseline for tuning our future performance for this task.

Quite a different picture appears in our “POST” runs, shown in Table 2. All of these runs use expansion of the queries based on the “TITLE” and “LOCATION” elements of the metadata annotations associated with images included in the “image” element of the queries (e.g. rather like image processing approaches, but using only text). All of these runs, had they been official runs, would have ranked at or near the top of the evaluation pools. (The similarity in between our MAP results and some others would seem to indicate that others were using a similar method of query expansion, perhaps in addition to image processing).

In Table 3 we compare the best performing runs, all using blind relevance feedback, for matching tasks in our official and unofficial runs. For Bilingual English to German we see a dramatic 146% improvement in MAP using this query expansion method. For the comparable monolingual tasks we see improvements ranging from almost 50% to over 80%. This is strong indication that expanding queries using the metadata of relevant images is a very good strategy for the ImageCLEFphoto task.

As a further experiment, we decided to try this expansion method for Bilingual retrieval *without any translation* of the original topic. That is, we used the title in the original language (in this case French) and did our expansion using appropriate metadata for the target language (English). The result from this experiment is included in Table 2 as run “POST_BLFREN_T_EXP-TL”. If that run had been submitted as an official run, it would have been a 186.5% improvement in MAP over the best official French to English submitted for the ImageCLEFphoto task.

References

1. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the imageclef 2006 photographic retrieval and object annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain. LNCS (September 2006)
2. Larson, R., Carson, C.: Information access for a digital library: Cheshire II and the Berkeley environmental digital library. In: Woods, L. (ed.) Knowledge: Creation, Organization and Use: Proceedings of the 62nd ASIS Annual Meeting, Medford, NJ, pp. 515–535. Information Today (1999)
3. Larson, R.R., Gey, F.C.: Geoclef text retrieval and manual expansion approaches. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain. LNCS (2006) (to appear)

Expanding Queries Through Word Sense Disambiguation^{*}

J.L. Martínez-Fernández¹, Ana M. García-Serrano², Julio Villena Román^{1,3},
and Paloma Martínez¹

¹ Universidad Carlos III de Madrid,
{joseluis.martinez, julio.villena, paloma.martinez}@uc3m.es

² Facultad de Informática, Universidad Politécnica de Madrid,
agarcia@dia.fi.upm.es

³ DAEDALUS - Data, Decisions and Language, S.A.
jvillena@daedalus.es

Abstract. The use of semantic information in the right way can lead to improved precision and recall in Information Retrieval (IR) systems. This assumption is the start point for the work carried out by the MIRACLE research team at ImageCLEF 2006. For this purpose, an implementation of the specification marks Word Sense Disambiguation (WSD) method **Error! Reference source not found.** has been developed. This method is based on WordNet **Error! Reference source not found.** and tries to select the right sense of each word appearing in the query. This allows the inclusion of only the correct synonyms when a semantic expansion is done. This selective expansion method has been combined with a deeper linguistic analysis to interpret negations and filter out common phrases and expressions used in query captions. Obtained results are included in Section 3 and, unfortunately, show that this approach to deal with semantics in IR systems does not lead to better precision and recall.

1 Introduction

This paper describes the image retrieval experiments developed for the ImageCLEF 2006 ad-hoc photographic retrieval task **Error! Reference source not found.** by the MIRACLE team. This team is made up of three research groups from three different universities located in Madrid (UPM, UC3M and UAM) along with a private company, DAEDALUS, the coordinator of the MIRACLE team. The team has taken part in CLEF since 2003 in many different tasks.

2 Semantic Expansion and Linguistic Analysis

Two new techniques have been developed for the 2006 ad-hoc image retrieval task: a

^{*} Work partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and by the Spanish R+D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01.

semantic expansion method based on WordNet and an interpreter to filter out common expressions and negation structures. Our previous experiences in image retrieval processes **Error! Reference source not found.** have shown that, in practical applications, current indexing techniques have reached their precision and recall upper limit. To surpass this point, we believe that semantic information should be considered in the retrieval process. For this reason the system developed this year includes an implementation of a WordNet based semantic expansion method that uses specification marks **Error! Reference source not found.** (adapted to WordNet 2.1) and a topic analysis module, which is intended to detect common words and to filter out words introduced by negation expressions.

The semantic expansion method was defined to disambiguate words appearing in WordNet, using the context of the word to select the correct sense among the set of senses assigned by WordNet. Intuitively, the idea is to select the sense pertaining to the hyperonym of the word that includes the greater number of senses for the words appearing in the context. The main objective is to include only the synonyms corresponding to the correct sense or the word instead of adding all synonyms for all senses of a word. Not every word can be disambiguated by applying this algorithm. Thus, to increase the number of correctly disambiguated words, three heuristics were identified. Only one of this heuristics has been included in our implementation, the *Definition Heuristic*, which discards senses whose glosses do not include any word present in the context. An example of the result of the expansion process applied to the title of topic 30 is shown in Fig. 1.

WordNet based Semantic Expansion method applied to the title of Topic 30: "room with more than two beds"		
Filtered nouns	Semantic Expansion without WSD	Semantic Expansion with WSD
beds, room	beds, furniture, "piece of furniture", "article of furniture", plot, "plot of ground", patch, bottom, "natural depression", depression, stratum, layer, sheet, "flat solid", surface, foundation, base, fundament, foot, groundwork, substructure, understructure, room, area, way, "elbow room", position, "spatial relation", opportunity, chance, gathering, assemblage	beds, furniture, "piece of furniture", "article of furniture", room, area

Fig. 1. Example of the semantic expansion method

On the other hand, a topic analysis module has been developed to make a deeper selection of terms to search against the index. This module is used to filter out common words and expressions as well as to detect negation structures. Thus, phrases like *“Relevant images will show”* are excluded and words accompanying phrases like *“are not relevant”* are expressed with a negation symbol “-“ which is interpreted by the search engine. When negations are interpreted, not every word taking part in the

sentence is excluded, only those words that do not appear in an affirmative form are marked as negative.

3 Description of Experiments

Different combinations of techniques lead to different experiments. Although the IAPR collection is available in two languages, English and German, only the English target language has been considered. On the other hand, four different query languages have been tested: English (EN), Japanese (JA), Polish (PL), Russian (RU) and Traditional Chinese (ZHT). Names and descriptions of runs are included in Table 1. The column “*TP*” (Topic Part) marks the field of the topic that has been processed: if the label “*T*” is included, it means that only the title fragment of the topic has been used, whereas if the label “*T+N*” is given, it means that both fields of the topic have

Table 1. MAP figures for text-based experiments

Run Name	TL	TP	TA	Exp.	Index	MAP
miratnntdenen	EN	T+ N	Noun	No	T+ D	0,1967
miranntdenen	EN	N	Noun	No	T+ D	0,1947
mirancdenen	EN	N	Common	No	T+ D	0,1875
miratnctdenen	EN	T+ N	Common	No	T+ D	0,1866
miratnndenen	EN	T+ N	Noun	No	D	0,1388
miratncdenen	EN	T+ N	Common	No	D	0,1375
miranndenen	EN	N	Noun	No	D	0,1371
mirancdenen	EN	N	Common	No	D	0,1314
miratntdjaen	JA	T	Noun	No	T+ D	0,1252
miratctdjaen	JA	T	Common	No	T+ D	0,1252
miratnndtddenen	EN	T+ N	Noun	Yes	T+ D	0,1176
miranndtddenen	EN	N	Noun	Yes	T+ D	0,1176
miratctdplen	PL	T	Common	No	T+ D	0,1075
miratndtdjaen	JA	T	Noun	Yes	T+ D	0,1049
miratntdzhsen	ZHT	T	Noun	No	T+ D	0,1041
miratctdzhsen	ZHT	T	Common	No	T+ D	0,1041
miratntdplen	PL	T	Noun	No	T+ D	0,101
miratctdrunen	RU	T	Common	No	T+ D	0,0909
miratndtdzhsen	ZHT	T	Noun	Yes	T+ D	0,0854
miratnnddenen	EN	T+ N	Noun	Yes	D	0,0837
mirannddenen	EN	N	Noun	Yes	D	0,0835
miratntdrunen	RU	T	Noun	Yes	T+ D	0,0795
miratndtdplen	PL	T	Noun	Yes	T+ D	0,0758
miratncdtddenen	EN	T+ N	Common	Yes	T+ D	0,074
mirancdtddenen	EN	N	Common	Yes	T+ D	0,0719

miratncddenen	EN	T+ N	Common	Yes	D	0,0583
mirancddenen	EN	N	Common	Yes	D	0,0578
miratndtenen	EN	T+ N	Noun	Yes	T	0,0525

been processed. Bilingual runs are always marked with the label “T” because the title is the only field provided in topic in languages distinct than English. Values for the “TA” (Topic Analysis) column can be “Common”, if common expressions and negations are processed, or “Noun”. The “Exp.”(Expansion) tells if the Semantic Expansion module has been used. Finally, the “Index” column points out which section of the image caption is used, the title (T) fragment, the description (D) fragment or both.

4 Conclusions

Experiments results sent to the workshop held in September, 2006, in Alicante, Spain, were incorrect. This was due to an implementation error in the semantic expansion method. Results shown in Table 1 were recalculated once these errors were fixed. As can be seen, no better MAP figures have been produced. Although experiments where the disambiguation process was involved have improved the results, MAP values have not been as good as expected. When semantic expansion is carried out, more documents are retrieved, but some of the relevant ones are lost. This can be due to the variation in the order of the documents in the results list introduced by the semantic expansion process. It must be taken into account that only the first one thousand documents are considered to obtain the MAP. On the other side, it is worth mentioning the decrease in MAP figures compared to previous campaigns, highlighting the influence of the change in the image collection used in the task.

The first step for future work will be to evaluate the semantic expansion method in a WSD task using a standard corpus such as SemCor. On the other hand, semantic expansion methods not based on WordNet are being studied to be applied in future campaigns.

References

- [1] Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks, in Evaluation of Multilingual and Multi-modal Information Retrieval. In: Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS (2006) (to appear)
- [2] Fellbaum, C.: Wordnet: An electronic lexical database. The MIT Press, Cambridge (1998)
- [3] Martínez-Fernández, J.L., García-Serrano, A., Villena, J., Méndez-Sáez, V.: MIRACLE approach to ImageCLEF 2004: merging textual and content-based Image Retrieval. In: Peters, C., et al. (eds.) CLEF 2004 proceedings. LNCS, vol. 3491, Springer, Heidelberg (2005)
- [4] Montoyo, A.: Método basado en marcas de especificidad para WSD. In: Proceedings of SEPLN 24 (2000)

Using Visual Linkages for Multilingual Image Retrieval

Masashi Inoue

National Institute of Informatics, Tokyo, Japan
m-inoue@nii.ac.jp

Abstract. The use of visual features in text-based ad-hoc image retrieval is challenging. Visual and textual information have so far been treated equally in terms of their properties and combined with weighting mechanisms for balancing their contributions to the ranking. The use of visual and textual information in a single retrieval system sometimes limits its applicability due to the lack of modularity. In this paper, we propose an image retrieval method that separates the usage of visual information from that of textual information. Visual clustering establishes linkages between images and the relationships are later used for the re-ranking. By applying clustering on visual features prior to the ranking, the main retrieval process becomes purely text-based. Experimental results on the ImageCLEFphoto ad-hoc task show this scheme is suitable for querying multilingual collections on some search topics.

1 Introduction

In the case of annotation-based multilingual image retrieval without any translation, the target collection is limited to the images annotated in the query language. Even if users are aware of that there are many relevant images annotated in different languages, many may not spare any of their time to explore them and instead may consider a portion of images annotated in one language as sufficient. Through automation, cross-language information retrieval (CLIR) techniques may be of help by expanding the range of images accessible on some search topics.

Text-based or annotation-based retrieval is thought to be the basis of image retrieval because associated text information plays an important role in assuring the effectiveness of retrieval. However, annotation-based approaches have limitations due to the lack of sufficient textual annotations. One reason for this insufficiency is the high cost of creating annotations. Also, for most nonspecialists, it is not easy to imagine the need for annotating images for future use. Therefore, the question lies in how we can supplement textual information in existing image collections. To alleviate the problem of text scarcity in image retrieval, we propose to use a “knowledge injection” framework. The definition of “knowledge” in this paper is that the entities used correspond to human conceptualization and their relationships are intuitively understandable, often in the form of qualitative values. In addition, the segregation of provided and acquired

Table 1. Comparison of knowledge injection and information integration frameworks

Type of entity	Automatically generated	Manually created
<i>Interpretable</i>	Extracted knowledge injection	Provided knowledge injection
<i>Non-interpretable</i>	Data integration	Not defined

knowledge is important. From the main retrieval system’s side, knowledge bases that are the result of manual effort are called “provided”. In contrast, information automatically extracted from the target collection or other information source is called “acquired”. Table 1 illustrates the relationship between knowledge injection and data integration, as well as the difference between provided and acquired knowledge. An example of provided knowledge is the WordNet thesaurus that injects the knowledge of relationships between words [1]. It has been applied to annotation-based image retrieval [2]. We investigate the use of acquired knowledge based on visual features of images in a target collection.

The use of visual features is conducted in the framework of a “find similar” task. This procedure can be executed in two ways. The first way involves searching for similar images after the initial result has been retrieved and sometimes makes use of user feedback. An on-line method for image retrieval involves clustering the retrieved results. For example, Chen et al. used a clustering method, which is originally developed for content-based image retrieval, in the annotation-based image retrieval framework, as post-processing after querying [3]. The second way involves finding similar image pairs or groups prior to querying. We investigate this second option in this paper. This choice is based on considerations of efficiency. The similarity calculation between images based on visual features is usually heavier than the one based on textual features. Thus, for image retrieval, it is sometimes desirable to conduct such computations off-line.

In the following sections, we introduce the idea of micro-clustering pre-processing to extract visual knowledge and describe the configuration of our retrieval models. We then show and analyze retrieval results for the Image-CLEFphoto collection. Finally, we conclude the paper.

2 System Description

2.1 Micro-clustering

Our retrieval strategy consists of three distinct steps. The first step is the knowledge extraction by using micro-clustering. Two types of clustering can be imagined. One is macro-clustering, or global partitioning, when the entire feature space is divided into sub-regions. For the document collection $\mathcal{D}$, in the macro-clustering, the set of clusters $\mathcal{C}_g = \{c_1, c_2, \dots, c_M\}$ are constructed where $\mathcal{D} = \bigcup_m c_m$ holds. Also, all documents belong to only one cluster. That is, if

$d_i \in c_m, \forall j \neq i, d_j \notin c_m$ and $\forall i, d_i \in c_m (m = 1, \dots, M)$. The other is micro-clustering, or local pairing, where nearby data points are linked so that they form a small group in a particular small region of the feature space. The set of clusters $\mathcal{C}_l = \{c_1, c_2, \dots, c_M\}$ are constructed and there is not any constraints as above. That is, each document does not have to be a cluster member and can be a member of more than one cluster. Usually the number of clusters M in micro-clustering is bigger than macro-clustering and the size of individual clusters are far smaller. In our system, micro-clustering was used to group images based on their visual similarities.

The concept of the micro-clustering has been developed to enhance the indexing of textual documents [4]. In [4], micro-clusters are treated as documents and used for classification. We used the same concept but with the features and similarity measure for visual information. Further, in this paper, the micro-clusters are regarded as the linkage information and used for re-ranking.

The process of clustering is as follows. First, visual features are extracted from all images. Simple color histograms are used. Since the images are provided in true color JPEG format, the histograms are created for the red (R), green (G), and blue (B) elements of the images. This results in three vectors for each image: $\mathbf{x}_r$, $\mathbf{x}_g$, and $\mathbf{x}_b$. The length of each vector, or the size of the histogram, $i = 256$. These vectors are concatenated and define a single feature matrix for each image: $X = [\mathbf{x}_r, \mathbf{x}_g, \mathbf{x}_b]$. Thus, the size of the feature matrix is i by j where $j = 3$.

The similarities between images are calculated using the above feature values. The similarity measure employed was the two-dimensional correlation coefficient r between the matrices. Assuming two matrices A and B , the correlation coefficient is given as

$$r = \frac{\sum_i \sum_j (A_{ij} - \bar{A})(B_{ij} - \bar{B})}{\sqrt{(\sum_i \sum_j (A_{ij} - \bar{A})^2)(\sum_i \sum_j (B_{ij} - \bar{B})^2)}} \quad (1)$$

where $\bar{A}$ and $\bar{B}$ are the mean values of A and B respectively.

Next, a threshold is set that determines which two or more images should belong to the same cluster. In other words, image pairs whose r score is larger than the threshold are considered identical during retrieval. At this stage, the threshold value is manually determined by inspecting the distribution of similarity scores so that relatively small numbers of images constitute clusters. Small clusters containing nearly identical images are preferred since visual similarity does not correspond to semantic similarity; however, visual identity often corresponds to the semantic identity.

2.2 Initial Retrieval

The novelty of our method in the ImageCLEF2006 is solely in the pre-processing of the retrieval. For the second step of the process, we used an existing search engine, the Lemur Toolkit¹. We used a unigram language-modeling algorithm

¹ <http://www.lemurproject.org/>

for building the document models, and the Kullback-Leibler divergence for the ranking. The document models were smoothed using Dirichlet prior [5].

2.3 Re-ranking

The third step is the injection of the linkage knowledge extracted in the first step. The ranked lists given by the retrieval engine are re-ranked by using the cluster information. The ranked list is searched from the top and when an image that belongs to a cluster is found, all other members of the cluster are given the same score as the highly ranked one. This process is continued until the number of images in the list exceeds the pre-specified number, which is 1000 in our study.

3 Experiments

3.1 Experimental Configuration

The details of the test collection used are given in [6]. There are 20,000 images annotated in English and in German. Instead of viewing the collection as a single bilingual collection, it is regarded as a collection of 20,000 English images and a collection of 20,000 German images. Each annotation has seven fields but only the title and description fields were used.

For the comparison, six runs were tested for the monolingual evaluation. The query languages were English, German, and Japanese. The collection languages were English and German. We applied query translation. The Systran machine translation (MT) system² was used. Because of lack of direct translation functionality between German and Japanese in the MT system, English was used as the pivot language when querying German collections using Japanese topics. That is, Japanese queries were first translated into English, and then the English queries were translated into German. The relationship between query and document languages and monolingual runs' names is summarized in Table 2.

In the table, names of runs are assigned according to the following rules. The first element *mcp* comes from the proposed method, micro-clustering pre-processing, and represents our group. The second element *bl* indicates that the baseline method was used. When the micro-clustering pre-processing was used, the value of the threshold is used for this element. For example, *09* denotes that the pairs with correlation coefficients greater than 0.9 form a cluster. The next element concerns the query language and the fields of the search topics. Runs using English queries with only title fields are marked as *eng_t*. Similarly, the next element is the collection language and the fields of the annotations. Runs using the German collection with title and description fields are marked by *ger_td*. When half of the English collection and half of the German collection are mixed together, the notation is *half_td*, as shown in Table 3. The last element is the configuration of the retrieval engine. The simple Kullback-Leibler divergence

² <http://babelfish.altavista.com/>

Table 2. Summary of runs on monolingual collections (run names and MAP scores)

Query	Document Language			
Language	English		German	
English	mcp.bl.eng_t.eng_td.skl_dir	0.1193	mcp.bl.eng_t.ger_td.skl_dir	0.0634
German	mcp.bl.ger_t.eng_td.skl_dir	0.1069	mcp.bl.ger_t.ger_td.skl_dir	0.0892
Japanese	mcp.bl.jpn_t.eng_td.skl_dir	0.0919	mcp.bl.jpn_t.ger_td.skl_dir	0.0316

Table 3. Summary of runs on the linguistically heterogeneous collection (run names and MAP scores)

Query	Half English	Half German	Document	Collection
Language	Without pre-processing		With pre-processing	
English	mcp.bl.eng_t.half_td.skl_dir	0.0838	mcp.09.eng_t.half_td.skl_dir	0.0586
German	mcp.bl.ger_t.half_td.skl_dir	0.0509	mcp.09.ger_t.half_td.skl_dir	0.0374

measure was used for ranking (*skl*) and Dirichlet prior used for smoothing (*dir*). All runs used the same configuration.

3.2 Results for Homogeneous Collections

In the baseline runs, the collection language is the determining factor of retrieval performance as shown in Table 2. Searching an English collection is better in any query language. Furthermore, the translated topics from German to English on the English collection worked better than mono-lingual German topics on the German collection. The results for Japanese topics on the German collection were poor, because of poor machine translation.

3.3 Results for Heterogeneous Collections

The advantage of the visual-similarity based pre-clustering becomes clear from the application of linguistically heterogeneous image collections. Therefore, a linguistically heterogeneous collection was constructed by taking 10,000 randomly chosen images from the English collection and the remaining 10,000 images from the German collection. There were no overlapping images. Both English and German queries without translation were tested on this single collection with micro-clustering. Table 3 shows the result. Because half of relevant images are now in annotated in another language, the MAP scores were worse than mono-lingual cases. It was observed that the pre-processing gave no improvement in terms of the mean average precision (MAP) scores.

When we looked at the results topic-by-topic, we found improvements for some topics, as shown in Figure 1. This suggests the need for topic analysis when applying micro-clustering pre-processing.

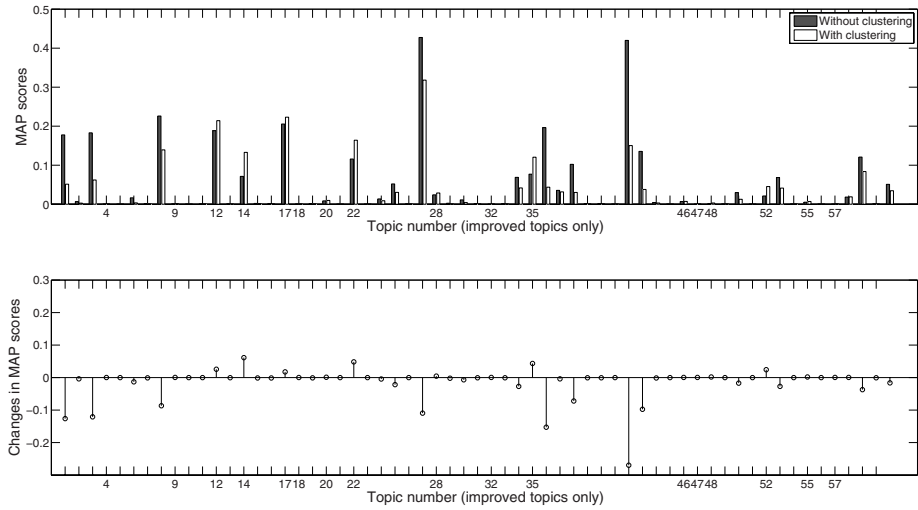


Fig. 1. MAP scores of each topic with and without pre-processing

3.4 Analysis of Clustering Result

The generated clusters were small and often of size two: a cluster formed by a pair of images. We intended this to be the result of micro-clustering; we wanted quite small yet highly condensed clusters. The statistics of cluster sizes are as follows: mean = 12.72, standard deviation = 43.81, minimum = 0, and maximum = 368. Some clusters have more than 100 members. Such non-microclusters are not ideal because when one of their members appears in the list, the cluster dominates the entire list after re-ranking. Thus, clusters bigger than 6 were truncated to size 6.

3.5 Discussion

Incorporating visual pre-processing did not improve the average performance for all topics. This failure might be because clusters of irrelevant images were used rather than relevant ones. Because not all of the initially retrieved images were relevant, we may need to use certain tactics to select only highly relevant images. Also, there is a trade-off between the quality of clustering and the degree of search target expansion. In the experiment, the threshold value might be conservative in order to avoid the inclusion of noisy clusters. More investigations are needed to clarify the effect of the threshold values.

The potential advantages of our approach over the usual query translation methods are as follows. First, there is no need to combine rankings given by multiple translated queries. Because the rank aggregation is difficult in IR, trial and error in the design of the merging strategies can not be avoided. Our approach outputs one ranking and the merging is not needed. Second, the systems

do not have to deal with the languages. The method can be used even when the language distribution within the collection is unknown.

The limitations of our experimental setting should be noted. The test collection is built upon a random selection from two language collections. Thus, nearly identical images that might have been originally created in a sequential manner could have been split into two languages. However, in reality, many similar image pairs may have annotations in only one language. For example, if one photographer took photos of an object, it would be natural to assume that all of these photos would be annotated in the same language. In the future, we will investigate more realistic linguistically heterogeneous collections.

In regard to the generalizability of the outcome of our experiment, the properties of the target collections is a big issue. The IAPR TC-12 collection used here mostly contains tourist photos. It contains a moderate amount of nearly identical images. If the target collection contained a larger amount of nearly identical images, the gain given by the pre-processing may be higher than it was with IAPR TC-12, whereas if it had few similar images, the preprocessing would not likely improve the retrieval effectiveness. Another issue is that the some images of a topic mostly exist in the sphere of one language. We do not have enough knowledge about the relationship between language and search topics yet.

4 Conclusion

This paper presented the results of experimental runs on the ImageCLEF2006 ad-hoc photo retrieval task. The goal was to investigate the possibility of a modular retrieval architecture that uses visual and textual information at different stages of retrieval. A visual feature-based micro-clustering was used for the linkage of nearly identical images annotated in different languages. After this pre-processing, the retrieval was conducted as a monolingual retrieval using query language. Then, images that are linked to the highly ranked images are pulled up. As a result, images annotated in different languages can be searched beyond the language barriers. In the experiment, although the mean performance over all topics did not improve, the individual average precision for some topics improved. The gains originated from the inclusion of additional images annotated in another language to the ranked list.

The biggest issue remaining is the lack of understanding of real-world needs for the cross-language image access. It is not fully known in what sort of search task that cross-language retrieval techniques will be helpful in information access in general [7]. Considering the language-independent nature of visual representation, cross-language image retrieval may be one such task. However, even in image retrieval, it is not clear how we can characterize the target collection from the perspective of image type and linguistic non-uniformity. Besides the refinement of pre-processing and post-processing, our future work will include developing a methodology for analyzing tasks and collections.

Acknowledgment

This research is partly supported by MEXT Grant-in-Aid for Scientific Research on Priority Areas (Cyber Infrastructure for the Information-explosion Era) and Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (17700166).

References

1. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* 38, 39–41 (1995)
2. Smeaton, A.F., Quigley, I.: Experiments on using semantic distances between words in image caption retrieval. In: *Proc. 19th ACM Int'l. Conf. on Research and Development in Information Retrieval*, Zurich, pp. 174–180 (1996)
3. Chen, Y., Wang, J.Z., Krovetz, R.: CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing* 14, 1187–1201 (2005)
4. Aizawa, A.: An approach to microscopic clustering of terms and documents. In: Ishizuka, M., Sattar, A. (eds.) *PRICAI 2002. LNCS (LNAI)*, vol. 2417, pp. 404–413. Springer, Heidelberg (2002)
5. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22, 179–214 (2004)
6. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the imageclef 2006 photographic retrieval and object annotation tasks. In: *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain. LNCS, Springer, Heidelberg (2006) (to appear)
7. Inoue, M.: The remarkable search topic-finding task to share success stories of cross-language information retrieval. In: *New Directions in Multilingual Information Access: A Workshop at SIGIR 2006*, Seattle, USA, pp. 61–64 (2006)

Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval

Yih-Chen Chang and Hsin-Hsi Chen*

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan

ycchang@nlg.csie.ntu.edu.tw, hhchen@csie.ntu.edu.tw

Abstract. Two kinds of intermedia are explored in ImageCLEFphoto2006. The approach of using a word-image ontology maps images to fundamental concepts in an ontology and measure the similarity between two images by using the kind-of relationship of the ontology. The approach of using an annotated image corpus maps images to texts describing concepts in the images, and the similarity of two images is measured by text counterparts using BM25. The official runs show that visual query and intermedia are useful. Comparing the runs using textual query only with the runs merging textual query and visual query, the latter improved 71%~119% of the performance of the former. Even in the situation which example images were removed from the image collection beforehand, the performance was still improved about 21%~43%.

1 Introduction

In recent years, many methods have been proposed to explore visual information to improve the performance of cross-language image retrieval (Clough, Sanderson, & Müller, 2005; Clough, et al., 2006). The challenging issue is the semantic differences among visual and textual information. For example, the visual information “red circle” may be ambiguous, i.e., images may containing “red flower”, “red balloon”, “red ball”, and so on, rather than the desired ones containing “sun”. That requires more contexts to retrieve the correct interpretation. The semantic difference between visual concept “red circle” and textual symbol “sun” is called a *semantic gap*.

Previous approaches have conducted text- and content-based image retrieval separately and then merged the results of two runs (Besançon, et al., 2005; Jones, et al., 2005; Lin, Chang & Chen, 2007). Content-based image retrieval may suffer from the semantic gap problem and report noise images. That may have negative effects on the final performance. Other approaches learn the relationships between visual and textual

* Corresponding author.

information and used the relationships for media transformation (Lin, Chang & Chen, 2007). The final retrieval depends on the performance of the relationship mining.

In this paper, we use two intermedia approaches to deal with the semantic gap. The main characteristic of these approaches is that human knowledge is imbedded in the intermedia and can be used to compute the semantic similarity of images. A word-image ontology and an annotated image corpus are explored and compared. Section 2 specifies how to build and use the word-image ontology. Section 3 deals with the uses of the annotated corpus. Sections 4 and 5 show and discuss the official runs in ImageCLEFphoto2006.

2 An Approach of Using a Word-Image Ontology

2.1 Building the Ontology

A word-image ontology is a word ontology aligned with the related images on each node. Building such an ontology manually is time consuming. In ImageCLEFphoto2006, the image collection has 20,000 colored images. There are 15,998 images containing English captions in <TITLE> and <DESCRIPTION> fields. The vocabularies include more than 8,000 different words, thus an ontology with only hundreds of words is not enough.

Instead of creating a new ontology from scratch, we extend WordNet, the well-known word ontology, to a word-image ontology. In WordNet, different senses and relations are defined for each word. For simplicity, we only consider the first two senses and kind-of relations in the ontology. Because our experiments in ImageCLEF2004 (Lin, Chang & Chen, 2006) showed that verbs and adjectives are less appropriate to be represented as visual features, we only used nouns here.

Before aligning images and words, we selected those nouns in both WordNet and image collection based on their TF-IDF scores. For each selected noun, we used Google image search to retrieval 60 images from the web. The returned images may encounter two problems: (1) they may have unrelated images, and (2) the related images may not be pure enough, i.e., the foci may be in the background or there may be some other things in the images. Zinger (2005) tried to deal with this problem by using visual features to cluster the retrieved images and filtering out those images not belonging to any clusters. Unlike Zinger (2005), we employed textual features. For each retrieved image, Google will return a short snippet. We filter out those images whose snippets do not exactly match the query terms. The basic idea is: “the more things a snippet mentions, the more complex the image is.” Finally, we get a word-image ontology with 11,723 images in 2,346 nodes.

2.2 Using the Ontology

2.2.1 Similarity Scoring

Each image contains several fundamental concepts specified in the word-image ontology. The similarity of two images is measured by the sum of the similarity of the fundamental concepts. In this paper we use kind-of relations to compute semantic

distance between fundamental concepts A and B in the word-image ontology. At first, we find the least common ancestor (LCA) of A and B . The distance between A and B is the length of the path from A through LCA to B . When computing the semantic distance of nodes A and B , the more the nodes should be traversed from A to B , the larger the distance is. In addition to the path length, we also consider the weighting of links in a path shown as follows.

(1) When computing the semantic distance of a node A and its child B , we consider the number of children of A . The more children A has, the larger the distance between A and B is. In an extreme case, if A has only one child B , then the distance between A and B is 0. Let $\#children(A)$ denote the number of children of A , and $base(A)$ denote the basic distance of A and its children. We define $base(A)$ to be $\log(\#children(A))$.

(2) When computing the semantic distance of a node A and its brother, we consider the level it locates. Assume B is a child of A . If A and B have the same number of brothers, then the distance between A and its brothers is larger than that between B and its brothers. Let $level(A)$ be the depth of node A . Assume the level of root is 0. The distance between node A and its child, denoted by $dist(A)$, is defined to be $C^{level(A)} \times base(A)$. Here C is a constant between 0 and 1. In this paper, C is set to 0.9.

If the shortest path between two different nodes N_0 and N_m is $N_0, N_1, \dots, N_{LCA}, \dots, N_{m-1}, N_m$, we define the distance between N_0 and N_m to be:

$$dist(N_0, N_m) = dist(N_{LCA}) + \sum_{i=1}^{m-1} dist(N_i) \quad (1)$$

The larger the distance of two nodes is, the less similar the nodes are.

2.2.2 Mapping into the Ontology

Before counting the semantic distance between two given images, we need to map the two images into nodes of the ontology. In other words, we have to find the fundamental concepts the two images consist of. A CBIR system is needed to do this. It regards an image as a visual query and retrieves the top k fundamental images (i.e., fundamental concepts) in the word-image ontology. In such a case, we have two sets of nodes $S_1 = \{n_{11}, n_{12}, n_{13}, \dots, n_{1k}\}$ and $S_2 = \{n_{21}, n_{22}, n_{23}, \dots, n_{2k}\}$, which correspond to the two images, respectively. We sum the CBIR scores of the k fundamental images with the given two images, respectively. The node with the highest score is selected as a basis, e.g., S_1 , to compute similarity. For each fundamental image n_{1i} ($1 \leq i \leq k$) for S_1 , we select the minimum distance between n_{1i} and n_{1j} ($1 \leq j \leq k$) of S_2 . The semantic distance between S_1 and S_2 is sum of the above k minimum distances.

Given a query with m example images, we regard each example image Q as an independent visual query, and compute the semantic distance between Q and images in the collection. Note that we determine what concepts are composed of an image in the collection before retrieval. After m image retrievals, each image in the collection has been assigned m scores based on the above formula. We choose the best score for each image and sort all the images to create a rank list. Finally, the top 1000 images in the rank list will be reported.

3 An Approach of Using an Annotated Image Corpus

An annotated image corpus is a collection of images along with their text annotations. The text annotation specifies the concepts and their relationships in the images. To measure the similarity of two images, we have to know how many common concepts there are in the two images. An annotated image corpus can be regarded as a reference corpus. We submit two images to be compared to the reference corpus using a CBIR system. The corresponding text annotations of the retrieved images are postulated to contain the concepts embedded in the two images. The similarity of text annotations measures the similarity of the two images indirectly.

The image collection in ImageCLEFphoto2006 can be considered as a reference annotated image corpus. Using image collection itself as intermedia has some advantages. It is not necessary to map images in the image collection to the intermedia. Besides, the domain can be more restricted to peregrine pictures. In the experiments, the <DESCRIPTION>, <NOTE>, and <LOCATION> fields in English form the annotated corpus.

To use the annotated image corpus as intermedia to compute similarity between example images and images in image collection, we need to map these images into intermedia. Since we use the image collection itself as intermedia, we only need to map example images in this work. An example image is considered as a visual query and submitted to retrieve images in intermedia by a CBIR system. The corresponding text counterparts of the top returned k images form a long text query and it is submitted to an Okapi system to retrieval images in the image collection. BM25 formula measures the similarity between example images and images in image collection.

4 Experiments

In the formal runs, we submitted 25 cross-lingual runs for eight different query languages. All the queries with different source languages were translated into target language (i.e., English) using Professional SYSTRAN system. We considered several issues, including (1) using different intermedia approaches (i.e., the text-image ontology and the annotated image corpus), and (2) with/without using visual query. In addition, we also submitted 4 monolingual runs which compared (1) the annotation in English and in German, and (2) using or not using visual query and intermedia. Finally, we submitted a run using visual query and intermedia only. The details of our runs are described as follows:

- (1) 8 cross-lingual and text query only runs:
 NTU-PT-EN-AUTO-NOFB-TXT, NTU-RU-EN-AUTO-NOFB-TXT,
 NTU-ES-EN-AUTO-NOFB-TXT, NTU-ZHT-EN-AUTO-NOFB-TXT,
 NTU-FR-EN-AUTO-NOFB-TXT, NTU-JA-EN-AUTO-NOFB-TXT,
 NTU-IT-EN-AUTO-NOFB-TXT, and NTU-ZHS-EN-AUTO-NOFB-TXT.

These runs are regarded as baselines and are compared with the runs using both textual and visual information.

- (2) 2 monolingual and text query only runs:
 NTU-EN-EN-AUTO-NOFB-TXT, and NTU-DE-DE-AUTO-NOFB-TXT.

These two runs serve as the baselines to compare with cross-lingual runs with text query only, and to compare with the runs using both textual and visual information.

- (3) 1 visual query only run with the approach of using an annotated image corpus: NTU-AUTO-FB-TXTIMG-WEprf.

This run will be merged with the runs using textual query only, and is also a baseline to compare with the runs using both visual and textual queries.

- (4) 8 cross-lingual runs, using both textual and visual queries with the approach of an annotated corpus:

NTU-PT-EN-AUTO-FB-TXTIMG-T-WEprf,
 NTU-RU-EN-AUTO-FB-TXTIMG-T-WEprf,
 NTU-ES-EN-AUTO-FB-TXTIMG-T-WEprf,
 NTU-FR-EN-AUTO-FB-TXTIMG-T-WEprf,
 NTU-ZHS-EN-AUTO-FB-TXTIMG-T-WEprf,
 NTU-JA-EN-AUTO-FB-TXTIMG-T-WEprf,
 NTU-ZHT-EN-AUTO-FB-TXTIMG-T-WEprf, and
 NTU-IT-EN-AUTO-FB-TXTIMG-T-WEprf.

These runs merge the textual query only runs in (1) and visual query only run in (3) with equal weight.

- (5) 8 cross-lingual runs, using both textual and visual queries with the approach of using word-image ontology:

NTU-PT-EN-AUTO-NOFB-TXTIMG-T-IOntology,
 NTU-RU-EN-AUTO-NOFB-TXTIMG-T-IOntology,
 NTU-ES-EN-AUTO-NOFB-TXTIMG-T-IOntology,
 NTU-FR-EN-AUTO-NOFB-TXTIMG-T-IOntology,
 NTU-ZHS-EN-AUTO-NOFB-TXTIMG-T-IOntology,
 NTU-JA-EN-AUTO-NOFB-TXTIMG-T-IOntology,
 NTU-ZHT-EN-AUTO-NOFB-TXTIMG-T-IOntology, and
 NTU-IT-EN-AUTO-NOFB-TXTIMG-T-IOntology.

These runs merge textual query only runs in (1), and visual query runs with weights 0.9 and 0.1.

- (6) 2 monolingual runs, using both textual and visual queries with the approach of an annotated corpus:

NTU-EN-EN-AUTO-FB-TXTIMG, and NTU-DE-DE-AUTO-FB-TXTIMG

These two runs using both textual and visual queries. The monolingual run in (2) and the visual run in (3) are merged with equal weight.

5 Results and Discussions

Table 1 shows experimental results of official runs in ImageCLEFphoto2006. We compare performance of the runs using textual query only, and the runs using both textual and visual queries (i.e., Text (T) Only vs. Text + Annotation (A) and Text + Ontology (O). In addition, we also compare the runs using word-image ontology and the runs using annotated image corpus (i.e., Text + Ontology vs. Text + Annotation). The runs whose performance is better than that of baseline (i.e., Text Only) will be marked in bold. The results show all runs using annotated image corpus are better than the baseline. In contrast, only two runs using word-image ontology are better.

Table 1. Performance of Official Runs

Query Language	MAP	Description	Runs
Portuguese	0.1630	T	NTU-PT-EN-AUTO-NOFB-TXT
	0.2854	T+A	NTU-PT-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1580	T+O	NTU-PT-EN-AUTO-NOFB-TXTIMG-T-IOntology
Russian	0.1630	T	NTU-RU-EN-AUTO-NOFB-TXT
	0.2789	T+A	NTU-RU-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1591	T+O	NTU-RU-EN-AUTO-NOFB-TXTIMG-T-IOntology
Spanish	0.1595	T	NTU-ES-EN-AUTO-NOFB-TXT
	0.2775	T+A	NTU-ES-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1554	T+O	NTU-ES-EN-AUTO-NOFB-TXTIMG-T-IOntology
French	0.1548	T	NTU-FR-EN-AUTO-NOFB-TXT
	0.2758	T+A	NTU-FR-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1525	T+O	NTU-FR-EN-AUTO-NOFB-TXTIMG-T-IOntology
Simplified Chinese	0.1248	T	NTU-ZHS-EN-AUTO-NOFB-TXT
	0.2715	T+A	NTU-ZHS-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1262	T+O	NTU-ZHS-EN-AUTO-NOFB-TXTIMG-T-IOntology
Japanese	0.1431	T	NTU-JA-EN-AUTO-NOFB-TXT
	0.2705	T+A	NTU-JA-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1396	T+O	NTU-JA-EN-AUTO-NOFB-TXTIMG-T-IOntology
Traditional Chinese	0.1228	T	NTU-ZHT-EN-AUTO-NOFB-TXT
	0.2700	T+A	NTU-ZHT-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1239	T+O	NTU-ZHT-EN-AUTO-NOFB-TXTIMG-T-IOntology
Italian	0.1340	T	NTU-IT-EN-AUTO-NOFB-TXT
	0.2616	T+A	NTU-IT-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1287	T+O	NTU-IT-EN-AUTO-NOFB-TXTIMG-T-IOntology

The reason why the word-image ontology does not perform as well as we expected may be that the images in the word-image ontology come from the web and the images in the web still contain much noise even after filtering. To deal with this problem, a better method of image filtering is necessary.

Since the example images in this task are in the image collection, the CBIR system always correctly maps the example images into themselves at mapping step. We made some extra experiments to examine the performance of our intermedia approach. In the experiments, we took out the example images from the image collection when mapping example images into intermedia. Table 2 shows the experiment results of the unofficial runs. Comparing Table 1 and Table 2 we find the performance of Table 2 is lower than that of Table 1. It shows the performance of CBIR in mapping stage will influence the final result and that is very critical. From Table 2, we also find that the approaches of annotated image corpus are better than the runs using textual query only. It shows even though there are some errors in mapping stage, the annotated image corpus can still work well.

Table 3 shows the experiment results of monolingual runs. Using both textual and visual queries are still better than runs using textual query only. The performance of the runs by taking out the example images from collection beforehand is still better than the runs use textual query only. From this table, we also find the runs using textual query only

Table 2. MAP of Unofficial Runs by Removing Example Images from the Collection

Query Language	Text Only	Text + Annotated image corpus
Portuguese	0.1630	0.1992
Russian	0.1630	0.1880
Spanish	0.1595	0.1928
French	0.1548	0.1848
Simplified Chinese	0.1248	0.1779
Japanese	0.1431	0.1702
Traditional Chinese	0.1228	0.1757
Italian	0.1340	0.1694

does not perform well even in monolingual runs. This may be because the image captions of this year are shorter and we do not have enough information when we use textual information only. In addition, when image captions are short too, the little differences in vocabularies between query and document may influence the results a lot. Therefore, German monolingual run and English monolingual runs perform very differently.

Table 4 shows the experiment of runs that use the visual query and annotated image corpus only, i.e., the textual query is not used. When example images were kept in the image collection, we can always map the example images into the right images. Therefore, the translation from visual information into textual information will be more correctly. The experiment shows the performance of visual query runs is better than that of textual query runs when the transformation is correct.

Table 3. Performance of Monolingual Image Retrieval

Query Language	MAP	Description	Runs
English	0.1787	T	NTU-EN-EN-AUTO-NOFB-TXT
(+example images)	0.2950	T+A	NTU-EN-EN-AUTO-FB-TXTIMG
(-example images)	0.2027	T+A	NTU-EN-EN-AUTO-FB-TXTIMG-NoE
German	0.1294	T	NTU-DE-DE-AUTO-NOFB-TXT
(+example images)	0.3109	T+A	NTU-DE-DE-AUTO-FB-TXTIMG
(-example images)	0.1608	T+A	NTU-DE-DE-AUTO-FB-TXTIMG-NoE

Table 4. Performance of Visual Query

MAP	Description	Runs
0.1787	T (monolingual)	NTU-EN-EN-AUTO-NOFB-TXT
0.2757	V+A (+example images)	NTU-AUTO-FB-TXTIMG-Weprf
0.1174	V+A (-example images)	NTU-AUTO-FB-TXTIMG-Weprf-NoE

6 Conclusion

The experiments show visual query and intermedia approaches are useful. Comparing the runs using textual query only with the runs merging textual query and visual query,

the latter improved 71%~119% of performance of the former. Even in the situation which example images are removed from the image collection, the performance can still be improved about 21%~43%. We find that the visual query in image retrieval is important. The performance of the runs using visual query only can be even better than the runs using textual only if we translate visual information into textual one correctly. The word-image ontology built automatically still contains much noise. We plan to investigate how to filter out the noise and explore different methods for cross-language image retrieval.

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC95-2221-E-002-334 and NSC95-2752-E-001-001-PAE.

References

1. Besançon, R., Héde, P., Moellic, P.A., Fluhr, C.: Cross-media feedback strategies: Merging text and image information to improve image retrieval. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 709–717. Springer, Heidelberg (2005)
2. Clough, P., Sanderson, M., Müller, H.: The CLEF 2004 cross language image retrieval track. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
3. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersch, W.: The CLEF 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
4. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 653–663. Springer, Heidelberg (2005)
5. Lin, W.C., Chang, Y.C., Chen, H.H.: Integrating textual and visual information for cross-language image retrieval: A trans-media dictionary approach. *Information Processing and Management, Special Issue on Asia Information Retrieval Research* 43(2), 488–502 (2007)
6. Zinger, S.: Extracting an ontology of portable objects from WordNet. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)

Dublin City University at CLEF 2006: Experiments for the ImageCLEF Photo Collection Standard Ad Hoc Task

Kieran McDonald* and Gareth J.F. Jones

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
{kmcdon,gjones}@computing.dcu.ie

Abstract. For the CLEF 2006 Cross Language Image Retrieval (ImageCLEF) Photo Collection Standard Ad Hoc task, DCU performed monolingual and cross language retrieval using photo annotations with and without feedback, and also a combined visual and text retrieval approach. Topics are translated into English using the Babelfish online machine translation system. Text runs used the BM25 algorithm, while visual approach used simple low-level features with matching based on the Jeffrey Divergence measure. Our results consistently indicate that the fusion of text and visual features is best for this task, and that performing feedback for text consistently improves on the baseline non-feedback BM25 text runs for all language pairs.

1 Introduction

Dublin City University's participation in the CLEF 2006 ImageCLEF Photo Collection Ad Hoc task adopted standard text retrieval with and without pseudo relevance feedback (PRF), and a combination of text retrieval with low-level visual feature matching. Text retrieval system is based on a standard Okapi model for document ranking and PRF [1]. Experiments are reported for monolingual English and German document retrieval and bilingual searching with a range of topic languages. Topics were translated for cross-language information retrieval using the online Babelfish machine translation engine [2]. Three sets of experiments are reported: first baseline text retrieval without PRF, second text retrieval with PRF, and finally the third set combines text retrieval and visual feature matching.

The results of our experiments demonstrate that PRF improves on the baseline in all cases with respect to both average precision and the number of relevant documents retrieved. Combined text retrieval with visual feature matching gives a further improvement in retrieval effectiveness in all cases.

This remainder of this paper is organised as follows: Section 2 briefly outlines the details of our standard retrieval system and describes our novel PRF method, Section 3 details our submitted runs, Section 4 gives results and analysis of our experiments, and finally Section 5 concludes the paper.

* Now at MSN Redmond, U.S.A.

2 System Description

The introduction of a new collection for the CLEF 2006 Ad Hoc Photo retrieval task meant that there were no existing retrieval test collections to use for system development. We thus used the previous ImageCLEF St Andrew's collection and related experiments on the TRECVID datasets to guide our selection of fusion methods and retrieval parameters for our experiments.

2.1 Text Retrieval

The contents of the structured annotation for each photo (TITLE, DESCRIPTION, NOTES, LOCATION and DATE fields) were collapsed into a flat document representation. Documents and search topics were processed to remove stopwords from the standard SMART list [3], and suffix stripped using the Snowball implementation of Porter stemming for the English language [4] [5]. While the German language topics and documents were stopped and stemmed using the Snowball German stemmer and stopword list [4].

Based on the development experiments, the text feature was matched using the BM25 algorithm with parameters: $k_1 = 1.0$, $k_2 = 0$, and $b = 0.5$. When using relevance feedback, the top 15 documents were assumed pseudo-relevant and the top scoring 10 expansion terms calculated using the Robertson selection value [1] were added to the original topic. The original query terms were upweighted by a factor of 3.5 compared to the feedback query expansion terms.

2.2 Visual Retrieval

The visual features were matched using the Jeffrey Divergence (a.k.a. Jensen-Shannon distance) matching function [6, 7]. We can interpret Jeffrey Divergence as measuring the efficiency of assuming that a common source generated both distributions – the query and the document.

The following three visual features were used: HSV colour histogram with 16x4x4 quantisation levels on a 5x5 regional image grid, Canny 8 edge + 1 bin for non-edges feature for each region of a 5x5 regional image grid, and a DCT histogram feature based on the first 5 coefficients quantised each into 3 values for a 3x3 regional image grid. More details on these features can be found in [8].

The results for the three visual features were combined using the weighted variant (i.e. linear interpolation) of the CombSUM fusion operator [9]. The scores for each feature were normalised between 0 and 1 and then the weighted sum was calculated for each document across the three features. The weights used for the colour, edge and DCT feature were respectively: 0.50, 0.30 and 0.20.

The results from the two visual examples for each topic were fused using the CombMAX fusion operator, which took the maximum of the normalised scores from each separate visual result list [9]. Scores were first normalised in the separate visual result sets for each topic image to lie between 0 and 1.

2.3 Text and Visual Retrieval Result Combination

Text and visual runs were fused using the weighted CombSUM operator with weights 0.70 and 0.30 for text and image respectively. Scores were normalised to lie between 0 and 1 in the separate text and visual result sets prior to fusion.

3 Description of Runs Submitted

We submitted two monolingual runs for German and English and cross language runs for both these languages. For English photo annotations we submitted runs for Russian, Portuguese, Dutch, Japanese, Italian, French, Spanish, German and Chinese topics. While for German photo annotations we only submitted runs for French and English topics. For each language pair including the monolingual runs we evaluated three different approaches: text only queries with and without PRF, and a combined visual and text (with text feedback) run for a total of 39 runs submitted.

4 Summary of Experimental Results

Results for our runs are shown in Table 1. From the table we can see that our multilingual fused text and visual submission performed very well, achieving either the second or top rank of submissions for each language pair in terms of Mean Average Precision (MAP). Text feedback increased the text only results in terms of MAP by on average 16.0%. Fusion with visual results increased these results on average by a further 9.2%. Our experiments produced consistent evidence across language pairs that text feedback and fusion with visual features is beneficial in Image Photo search.

Our monolingual English runs with text feedback and fused visual results performed relatively poorly, achieving 8th of 50 submissions in terms of MAP. The increased competition in the more popular monolingual English category relative to the multilingual categories as well as our limited approach in tuning a single parameter selection for all our submitted runs probably accounts for the relative decrease in effectiveness of our approach in the English monolingual category. Our system parameters were tuned for multilingual topics and in future we should tune separately for the monolingual case. For monolingual searches we would expect the effectiveness of the initial query to be higher than in the multilingual case, and therefore both the appropriate feedback and fusion parameters for combining text and visual results may differ significantly compared to the multilingual case. In our case, the initial English monolingual text query under-performed, but was increased by 20.3% by the text feedback approach we employed. This result was further improved by only 4% through fusion with visual results.

Our results consistently show that our fused text and image retrieval submission outperforms our text-only methods. The average relative improvement in MAP was 9.2%, maximum was 18.9% and minimum was 4.0%. The evaluation

Table 1. Retrieval results for our submitted monolingual and multilingual runs for the ImageCLEF Photo 2006 task. Each run is labelled by its Query-Document language pair, whether it used text feedback (FB) and the mediums fused: text only (TXT) or text fused with image search (TXT+IMG). Each run’s effectiveness is tabulated with the evaluation measures: Mean Average Precision (MAP), precision at document cut-offs 10 (P@10), 20 (P@20) and 30 (P@30), the number of relevant documents retrieved for each run (Rel. Ret) and the ranking of the results in terms of MAP compared to all other submitted runs to ImageCLEF Photo 2006 task for the respective language pair. The relative increase in MAP (%chg) between text with feedback and without, and between fused visual and text to text alone (i.e. text with feedback) is also listed for each run.

<i>Lang-Pair</i>	<i>FB</i>	<i>Media</i>	<i>MAP</i>	<i>%chg. P@10</i>	<i>P@20</i>	<i>P@30</i>	<i>Rel. Ret.</i>	<i>Rank</i>
MONOLINGUAL RUNS								
EN-EN	FB	TXT+IMG	.2234	4.0%	.3067	.2792 .2628	2205	8/50
	FB	TXT	.2148	20.3%	.3050	.2733 .2467	2052	10/50
	.	TXT	.1786	.	.2550	.2275 .2100	1806	21/50
DE-DE	FB	TXT+IMG	.1728	7.4%	.2283	.2425 .2328	1812	2/8
	FB	TXT	.1609	12.7%	.2283	.2300 .2083	1478	3/8
	.	TXT	.1428	.	.2283	.2050 .2017	1212	4/8
MULTILINGUAL RUNS WITH GERMAN DOCUMENTS								
EN-DE	FB	TXT+IMG	.1219	14.8%	.1850	.1750 .1694	1553	1/5
	FB	TXT	.1062	18.9%	.1400↓	.1358 .1272	1121	2/5
	.	TXT	.0893	.	.1450	.1317 .1211	814	4/5
FR-DE	FB	TXT+IMG	.1037	18.9%	.1517	.1467 .1461	1605	1/3
	FB	TXT	.0872	28.8%	.1317	.1225 .1167	1192	2/3
	.	TXT	.0677	.	.1150	.1083 .0994	898	3/3
MULTILINGUAL RUNS WITH ENGLISH DOCUMENTS								
CH-EN	FB	TXT+IMG	.1614	10.6%	.2267	.2042 .1939	1835	2/10
	FB	TXT	.1459	17.0%	.2033	.1908 .1706	1574	3/10
	.	TXT	.1247	.	.1783	.1650 .1606	1321	6/10
DE-EN	FB	TXT+IMG	.1887	5.8%	.2600	.2575 .2411	2014	1/8
	FB	TXT	.1783	10.5%	.2450	.2342 .2117	1807	2/8
	.	TXT	.1614	.	.2233	.2025 .1894	1642	3/8
ES-EN	FB	TXT+IMG	.2009	6.0%	.2633	.2525 .2467	2118	2/6
	FB	TXT	.1895	17.6%	.2583	.2367 .2117	1940	3/6
	.	TXT	.1612	.	.2183	.2075 .1978	1719	4/6
FR-EN	FB	TXT+IMG	.1958	6.5%	.2483	.2558 .2489	2026	2/7
	FB	TXT	.1838	13.3%	.2367	.2308 .2150	1806	3/7
	.	TXT	.1622	.	.2167	.2250 .1972	1652	4/7
IT-EN	FB	TXT+IMG	.1720	12.8%	.2167	.2200 .2150	2017	2/14
	FB	TXT	.1525	15.2%	.2150	.1833 .1711	1780	3/14
	.	TXT	.1324	.	.1917	.1650 .1556	1602	9/14
JP-EN	FB	TXT+IMG	.1615	10.7%	.2267	.2042 .1939	1848	2/10
	FB	TXT	.1459	17.0%	.2033	.1908 .1706	1591	3/10
	.	TXT	.1247	.	.1783	.1650 .1606	1321	8/10
NL-EN	FB	TXT+IMG	.1842	6.7%	.2467	.2342 .2194	1906	1/3
	FB	TXT	.1726	12.5%	.2150	.2042 .1900	1665	2/3
	.	TXT	.1534	.	.1817	.1775 .1706	1477	3/3
PT-EN	FB	TXT+IMG	.1990	7.8%	.2683	.2625 .2478	2032	2/6
	FB	TXT	.1846	13.8%	.2683	.2525 .2228	1824	3/6
	.	TXT	.1622	.	.2483	.2217 .1956	1642	5/6
RU-EN	FB	TXT+IMG	.1816	7.3%	.2600	.2400 .2156	1982	2/8
	FB	TXT	.1693	9.9%	.2500	.2158 .1911	1760	3/8
	.	TXT	.1541	.	.2333	.1867 .1694	1609	6/8

measures are improved for all tested language pairs. This indicates that our fusion approach is stable and produces reliable results. We suspect that our fusion parameters were a bit conservative in the importance given to the visual results for this task. But on the other hand, if we increase the importance of visual results, we may sacrifice some of the stability and consistency of our results. This will be investigated in followup experiments.

Our results also consistently show that our text runs are improved for all language pairs tested when using text feedback compared to without it. This is true for all language pairs and evaluation measures in Table 1, except for precision at a cut-off of 10 documents for English queries against German documents. The decrease in precision is small and insignificant in this case. At an average relative increase in MAP of 16.0%, a maximum of 28.8% and a minimum of 9.9% for the language pairs, the importance of text feedback is well established for text-based search in this task.

The lack of relevant tuning data because of the significant difference between this year and previous years ImageCLEF photo content and descriptions may have led to a less than optimal choice of parameter for fusing visual and text results. Post-ImageCLEF experiments should be able to quantify the improvements that can be made with better tuning or alternative fusion strategies.

5 Conclusions

Our results for the ImageCLEF 2006 photo retrieval task show that fusing text and visual results achieves better effectiveness than text alone. We also demonstrated that PRF is important for improving the effectiveness of the text retrieval model with consistent improvement in results across language pairs. Future experiments will investigate fusion of text and visual features more deeply, since we believe this still has more to offer than we have shown in our current experiments.

References

- [1] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In: Harman, D.K. (ed.) *Proceedings of the Third Text REtrieval Conference (TREC-3)*, NIST, pp. 109–126 (1995)
- [2] Babelfish, <http://babelfish.altavista.com/>
- [3] SMART, <ftp://ftp.cs.cornell.edu/pub/smart/>
- [4] Snowball toolkit <http://snowball.tartarus.org/>
- [5] Porter, M.F.: An algorithm for suffix stripping. *Program* 14, 10–137 (1980)
- [6] Rao, C.R.: Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya: The Indian Journal of Statistics* 44(A), 1–22 (1982)
- [7] Liu, J.: Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37(1), 145–151 (1991)
- [8] McDonald, K.: *Discrete Language Models for Video Retrieval*, Ph.D. Thesis, Dublin City University (2005)
- [9] Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: *Proceedings of the Third Text REtrieval Conference (TREC-1994)*, pp. 243–252 (1994)

Image Classification with a Frequency–Based Information Retrieval Scheme for ImageCLEFmed 2006

Henning Müller¹, Tobias Gass², and Antoine Geissbuhler¹

¹ Medical Informatics, University and Hospitals of Geneva, Switzerland
henning.mueller@sim.hcuge.ch

² Lehrstuhl für Informatik 6, RWTH Aachen, Germany
gass@informatik.rwth-aachen.de

Abstract. This article describes the participation of the University and Hospitals of Geneva at the ImageCLEF 2006 image classification tasks (medical and non–medical). The techniques applied are based on classical tf/idf weightings of visual features as used in the GIFT (GNU Image Finding Tool). Based on the training data, features appearing in images of the same class are weighted higher than features appearing across classes. These feature weights are added to the classical weights. Several weightings and learning approaches are applied as well as quantisations of the features space with respect to grey levels. A surprisingly small number of grey levels leads to best results. Learning can improve the results only slightly and does not obtain as good results as classical image classification approaches. A combination of several classifiers leads to best final results, showing that the schemes have independent results.

Keywords: Image Retrieval, Classification, Frequency-based.

1 Introduction

ImageCLEF¹ makes available realistic test collections for the evaluation of retrieval and classification tasks in the context of CLEF² (Cross Language Evaluation Forum). A detailed description of the object annotation task and a photographic retrieval task can be found in [1]. The overview includes a description of the tasks, submitted results, and a ranking of the best systems. A description of a medical image retrieval and automatic image annotation task can be found in [2] with all the details of submissions. More on the data can also be found on ³.

This article focuses on the submission for the two image classification tasks. The submissions were slightly after the deadline because of a lack of man power but can be compared with the results in the overview articles. Already in 2005, an automatic medical image annotation task was offered in ImageCLEF [3]. Best

¹ <http://ir.shef.ac.uk/imageclef/>

² <http://www.clef-campaign.org/>

³ <http://ir.ohsu.edu/image/>

results were obtained by systems using classical image classification techniques [4]. Approaches based on information retrieval techniques [5] had lower results but were still among the best five groups, without using any learning data. It was expected that a proper use of learning data could improve results significantly. Such a learning approach is attempted in this paper.

2 Methods

The methods described in this paper rely on the GIFT⁴ (GNU Image Finding Tool) [6]. The learning approaches are based on [7].

2.1 Features

GIFT uses four features groups described in more detail in [6].

- Global color histogram based on the HSV (Hue, Saturation, Value) quantised into 18 hues, 3 saturations, 3 values and 4 grey levels.
- Local color blocks. Each image is recursively partitioned into 4 blocks of equal size, and each block is represented by its mode color.
- A global texture histogram of the responses to Gabor filters of 3 scales and 4 directions, quantised into 10 bins.
- Local Gabor blocks by applying the filters above to the smallest blocks created by the recursive partition and using the same quantisation.

This results in 84'362 possible features where each image contains around 1'500. During the experiments, we increased the number of grey levels used for the color block features and color histogram features to 8, 16 and 32.

2.2 Feature Weights

The basic weighting used is *term frequency/inverted document frequency* (*tf/idf*), which is well known from text retrieval. Given a query image q and a possible result image k , a score is calculated as the sum of all weights .

$$\text{score}_{kq} = \sum_j (\text{feature weight}_j) \quad (1)$$

The weight of each feature is computed as follows using term frequency(*tf*) and collection frequency(*cf*):

$$\text{feature weight}_j = tj_j * \log^2(1/(cf_j)) \quad (2)$$

This results in giving frequent features a lower weight. We use the approach in [7] and add learning strategies to optimise results for classification.

⁴ <http://www.gnu.org/software/gift/>

Strategies. The learning approach commonly uses log files and finds *pairs* of images marked together. Frequencies can be computed of how often each feature occurs in a pair. A weight can be calculated using the information whether the images in the pair were both marked as *relevant* or whether one was marked *relevant* and the other as *notrelevant*. This results in desired and non-desired cooccurrence of features. In this paper, we want to train weights focused on classification. This means that we look at class memberships of images. Each result image is marked as relevant if the class matches that of the query image and non-relevant otherwise. We then applied several strategies for extracting the pairs of images. First, each pair of images occurring at least once is considered relevant. In the second approach we aim at discriminating positive and negative results more directly. Only the best positive and worst negative results of a query are taken into account. In a third approach, we pruned all queries which seem *too easy*. If the first N results were already positive we omitted the entire query from further evaluation. This is based on ideas similar to Support Vector Machines, where only information on class boundaries is taken into account.

Computation of Additional Feature Weights. For each image pair, we calculate the features they have in common and whether these were positive or negative. We used two ways to compute an additional factor:

- Basic Frequency : Features are weighted by the number of occurrences in positive pairs, normalised by the number of all pairs.

$$\text{factor}_j = \frac{|\{f_j | f_j \in I_a \wedge f_j \in I_b \wedge (I_a \rightarrow I_b)_+\}|}{|\{f_j | f_j \in I_a \wedge f_j \in I_b \wedge ((I_a \rightarrow I_b)_+ \vee (I_a \rightarrow I_b)_-)\}|} \quad (3)$$

where f_j is a feature j , I_a and I_b are two images and $(I_a \rightarrow I_b)_{+/-}$ denotes that I_a and I_b were marked together positively (+) or negatively (-).

- Weighted Probabilistic :

$$\text{factor}_j = 1 + (2 * \frac{pp}{|\{(I_a \rightarrow I_b)_+\}|}) - \frac{np}{|\{(I_a \rightarrow I_b)_-\}|} \quad (4)$$

where pp (positive probability) is the probability that feature j is important, whereas np (negative probability) denotes the opposite.

The additional factors calculated in this way are then simply multiplied with the already existing feature weights for the calculation of similarity scores.

2.3 Classification

For each query image q , a set of $N \in \{1, 3, 5, 10\}$ result images k with a similarity score S_k was returned. The class of each result image was computed and the similarity scores were added up for the corresponding classes. The class with the highest accumulated score was assigned to the image. From preliminary experiments it was visible that $N = 5$ produced the best results. This is similar to a typical K-nearest neighbour (k-NN) classifier.

3 Results

3.1 Classification on the LTU Database

The non-medical automatic annotation task consisted of 14'035 training images from 21 classes. Subsets of images such as *computer equipment* were formed, mainly with images crawled from the web with a large variety. The task was hard and only four groups participated. The content of the images was regarded as extremely heterogeneous even for same classes. Without using any learning methods, using a simple 5-NN classifier, GIFT had an error rate of 91,7%. Using the learning method with best/worst pruning and the frequency-based weighting, the error rate decreased to 90,5%. Best results when separately weighting the four feature groups were 88.3%

3.2 Classification on the IRMA Database

10'000 grey level images from 117 classes were made available as training data and 1'000 images as test data. Baseline results of GIFT with various quantisations of grey levels are in Table 1. They show clearly that more grey levels do not help classification, as error rates increase surprisingly.

Table 1. Error rates on the IRMA database using a varying number of grey levels

Number of grey levels	Error rate
4	32,0%
8	32,1%
16	34,9%
32	37,8%

Table 3.2 shows results of GIFT using learning approaches. Surprisingly, the effect of learning is small. The only method which improved the error rate at all was the tf/idf weighting combined with best/worst pruning.

Table 2. Error rates on the IRMA database using 4 grey levels

naive strategy	35,3%	32,4%
use best and worst results	31,7%	32,2%
removing too-easy queries	33,2%	32,5%

We combined eight grey levels with techniques but results were slightly worse. Finally, we accumulated scores of all runs performed resulting in an error rate of 29,7%, which shows that the approaches are combinable/independent.

4 Conclusion and Future Work

The provided tasks proved difficult to optimise for a frequency-based image retrieval system such as the GIFT using very simple features. Thus the features seem to be the main point for potential improvements when using a system similar to the GIFT. The various tested techniques showed that the system can profit from the training but that it needs to be done with much care. It can also be shown that tf/idf works very well on large collections without any knowledge but that classification, particularly when class sizes are very unbalanced needs more than lower weighting frequent features for good results. Some very frequent features might be important for the very large classes.

For future work it seems important to study features and feature groups independently as they are related and not independent. This means that a varying number of grey levels might be useful for local and global color features and that variations in the Gabor filters might also be useful in various ways (not discussed in this paper). Pre-treating images (background removal, normalisation of grey levels) allowing for more variation of the images with respect to object size and position are other approaches that are expected to improve results.

Acknowledgements

This work was partially supported by the Swiss National Science Foundation (Grant 205321-109304/1).

References

1. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the Image CLEF 2006 photo retrieval and object annotation tasks. In: Proceedings of CLEF 2006. LNCS, Springer, Heidelberg (to appear, 2007)
2. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Eugene, K., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In: Proceedings of CLEF 2006. LNCS, Springer, Heidelberg (2007)
3. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in ImageCLEF 2005: Combining content-based image retrieval with textual information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
5. Müller, H., Geissbuhler, A., Marty, J., Lovis, C., Ruch, P.: The use of MedGIFT and EasyIR for ImageCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)

6. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content-based query of image databases: inspirations from text retrieval. In: Ersboll, B.K., Johansen, P. (eds.) *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99*, vol. 21, pp. 1193–1198 (2000)
7. Müller, H., Squire, D.M., Pun, T.: Learning from user behavior in image retrieval: Application of the market basket analysis. *International Journal of Computer Vision* 56(1–2), 65–77 (2004) (Special Issue on Content-Based Image Retrieval)

Grayscale Radiograph Annotation Using Local Relational Features

Lokesh Setia, Alexandra Teynor, Alaa Halawani, and Hans Burkhardt

Albert-Ludwigs-University Freiburg
79110 Freiburg im Breisgau, Germany
{setia, teynor, halawani, burkhardt}@informatik.uni-freiburg.de

Abstract. Content based automatic image classification systems are increasingly finding usage, e.g. in large medical image databases. This paper concentrates on a grayscale radiograph annotation task which was a part of the ImageCLEF 2006. We use local features calculated around interest points, which have recently received excellent results for various image recognition and classification tasks. We propose the use of relational features, which are highly robust to illumination changes, and thus quite suitable for X-Ray images. Results with various feature and classifier settings are reported. A significant improvement in results is seen when the relative positions of the interest points are also taken into account during matching. For the given test set, our best run had a classification error rate of 16.7 %, just 0.5 % higher than the best overall submission, and therewith was ranked second in the medical automatic annotation task at the ImageCLEF 2006. The proposed method is general, can be applied to other image classification tasks and can also be extended to colour images.

Keywords: Local Features, Radiograph, Image Annotation, Invariants.

1 Introduction

Growing image collections of various kinds have led to the need for sophisticated image analysis algorithms to aid in classification and recognition. This is especially true for medical image collections, where the cost for manual classification of images collected through daily clinical routine can be overwhelming. Automatic classification of these images would spare the skilled human labour from this repetitive, error-prone task. As an example, a study carried out by the University Hospital in Aachen [1] revealed over 15% human errors in assignment of a specific tag, for X-ray images taken during normal clinical routine. Unless corrected, these images cannot be found again with keyword search alone. Our group (LMB) participated in the ImageCLEF 2006 medical automatic annotation task. In this paper, we describe our experiments and submitted runs which are performed using a local version of relational features which are very robust to illumination changes, and with different kinds of accumulators to aid in the

matching process. We conclude with a discussion of the positive and negative aspects of the various approaches.

2 Database

The database for the task was made available by the IRMA group from the University Hospital, Aachen, Germany. It consists of 10,000 fully classified radiographs taken randomly from medical routine. 1,000 radiographs for which classification labels were not available to the participants had to be classified. The aim is to find out how well current techniques can identify image modality, body orientation, body region, and biological system examined based on the images. The results of the classification step can be used for multilingual image annotations as well as for DICOM header corrections. The images are annotated with the complete IRMA code, a multi-axial code for image annotation. However, for the task, we used only a flat classification scheme with a total of 116 classes. Figure 1 shows that a great deal of intra-class variability exists, mostly in illumination changes, small amounts of position and scale differences, and noise. On the other hand, Figure 2 shows that the inter-class distance can be quite tight, as the four frontal chest categories are at least pairwise very similar to each other. The confusion matrix between these four classes indicates that more than 80 % of the images were correctly classified (with the Orientation Variant Accumulator described in Section 3.2. For a complete description of the database and the task, please refer to [7].

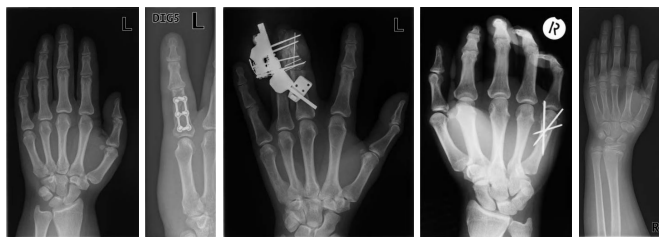


Fig. 1. Intra-class variability within the class annotated as “*x-ray, plain radiography, coronal, upper extremity (arm), hand, musculoskeletal system*”

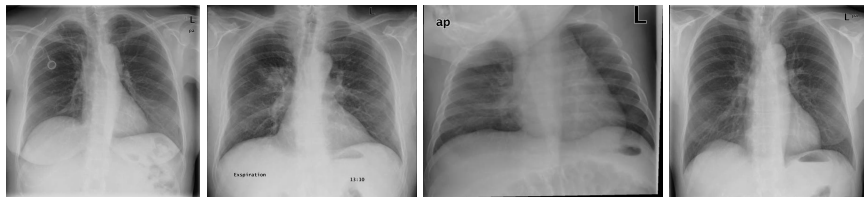


Fig. 2. One example image each from class number 108, 109, 110 and 111 respectively. Some classes are hard to distinguish by the untrained eye, but surprisingly perform quite well by automatic methods.

3 Features

3.1 Extraction of Local Features

We use features inspired from invariant features based on group integration. Let G be the transformation group for the transformations against which the invariance is desired. For the case of radiograph images this would be the group of translations. For rotation, only partial invariance is desired as most radiographs are scanned upright. If an element $g \in G$ acting on the images, the transformed image is denoted by $g\mathbf{M}$. An invariant feature must satisfy $F(g\mathbf{M}) = F(\mathbf{M}), \forall g \in G$. Such invariant features can be constructed by integrating $f(g\mathbf{M})$ over the transformation group G [2].

$$I(\mathbf{M}) = 1/|G| \int_G f(g\mathbf{M}) dg \quad (1)$$

Although the above features are invariant, they are not particularly discriminative, due to the fact that the integration runs over the whole image, which contains large insignificant regions. An intuitive extension is to use local features, which have received good results in several image processing tasks. The idea is to first use a so-called interest point detector, to select pixels with high information content in their neighbourhood. The meaning of “high information” differs from application to application. There are, for example, detectors which select points of high gradient magnitude, or corner points (points with multiple gradient orientations). A desirable characteristic expected from an interest-point detector is that it is robust towards the above mentioned transformations. In this work, we use the wavelet-based salient point detector proposed by Loupiau and Sebe [3].

For each salient point $\mathbf{s} = (x, y)$, we extract relational feature subvector around it, given as:

$$R_k = [\Re(x, y, r_1, r_2, \phi, n)]_k, \quad k = 1, \dots, n. \quad (2)$$

Then,

$$R_k = \text{rel}(\mathbf{I}(x_2, y_2) - \mathbf{I}(x_1, y_1)), \quad (3)$$

where

$$(x_1, y_1) = (x + r_1 \cos(k \cdot 2\pi/n), y + r_1 \sin(k \cdot 2\pi/n)), \quad (4)$$

and

$$(x_2, y_2) = (x + r_2 \cos(k \cdot 2\pi/n + \phi), y + r_2 \sin(k \cdot 2\pi/n + \phi)) \quad (5)$$

with the **rel** operator defined as:

$$\text{rel}(x) = \begin{cases} 1 & \text{if } x < -\epsilon \\ \frac{\epsilon-x}{2\epsilon} & \text{if } -\epsilon \leq x \leq \epsilon \\ 0 & \text{if } \epsilon < x \end{cases}, \quad (6)$$

The relational functions were first introduced for texture analysis by Schael [4], motivated by the local binary patterns [5]. For each (r_1, r_2, ϕ) combination a

local subvector is generated. In this work, we use 3 sets of parameters, $(0, 5, 0)$, $(3, 6, \pi/2)$ and $(2, 3, \pi)$, each with $n = 12$ to capture local information at different scales and orientations. The subvectors are concatenated to give a local feature vector for each salient point. The local feature vectors of all images are taken together and clustered using the k-means clustering algorithm. The number of clusters would be denoted here by N_c , and is determined empirically.

3.2 Building the Global Feature Vector

An equally important step is that of estimating similarity between two images, once local features have been calculated for both of them. In the scenario where the variability between the training and test images is strictly limited, correspondence-based matching algorithms can be quite effective, i.e. for each test local feature vector, the best training vector and its location is recalled (e.g. through nearest neighbour). The location information can be put to use by doing a consistency check if the constellation of salient points build a valid model for the given class.

However, the large intra-class variability displayed by the radiograph images suggests that a global feature based matching would perform better, as small number of non-correspondences would not penalize the matching process heavily. We build the following accumulator features from the cluster index image $I(\mathbf{s})$, i.e. an image which contains for each salient point, its assigned cluster index only.

All-Invariant Accumulator. The simplest histogram that can be built from a cluster index image is a 1-D accumulator counting the number of times a cluster occurred for a given image. All spatial information regarding the interest point is lost in the process. The feature can be mathematically written as:

$$\mathbf{F}(c) = \#\{\mathbf{s} \mid (I(\mathbf{s}) = c)\}, \quad c = 1, \dots, N_c$$

In our experiments, the crossvalidation results were never better than 60 %. It is clear that critical spatial information is being lost in the process of histogram construction.

Rotation Invariant Accumulator. We first incorporate spatial information, by counting pairs of salient points lying within a certain distance range, and possessing particular cluster indices, i.e. a 3-D accumulator defined by:

$$\mathbf{F}(c_1, c_2, d) = \#\{(\mathbf{s}_1, \mathbf{s}_2) \mid (I(\mathbf{s}_1) = c_1) \wedge (I(\mathbf{s}_2) = c_2) \\ \wedge (D_d < \|\mathbf{s}_1 - \mathbf{s}_2\|_2 < D_{d+1})\}$$

where the extra dimension d runs from 1 to N_d (number of distance bins).

The crossvalidation results improve to about 68 %, but it should be noted that this accumulator is rotation invariant (depends only on distance between salient points), while the images are upright. Incorporating unnecessary invariance leads

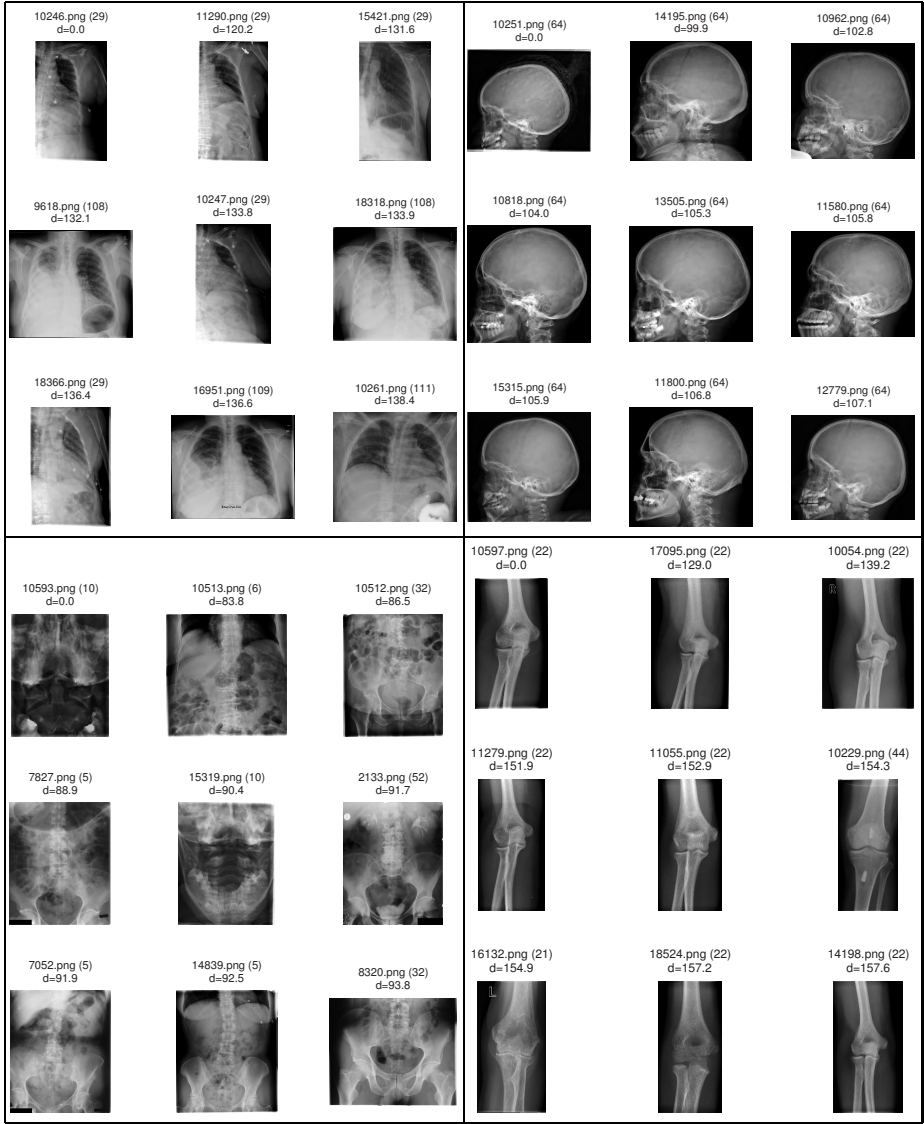


Fig. 3. Top 8 nearest neighbour results for different query radiographs. The top left image in each box is the query image, results follow from left to right, then top to bottom. The image caption contains the image name and label in (brackets). Finally the distance according to the L_1 norm is displayed underneath.

to a loss of discriminative performance. Especially in this task of radiograph classification it can be seen that mirroring the position of interest points leads to a completely another class, for example, a left hand becomes a right hand, and so on. Thus we incorporate orientation information in the next section.

Orientation Variant Accumulator. We discussed this method in detail in [6]. The following cooccurrence matrix captures the statistical properties of the joint distribution of cluster membership indices which were derived from local features.

$$\mathbf{F}(c_1, c_2, d, a) = \# \{ (\mathbf{s}_1, \mathbf{s}_2) \mid (I(\mathbf{s}_1) = c_1) \wedge (I(\mathbf{s}_2) = c_2) \\ \wedge (D_d < \|\mathbf{s}_1 - \mathbf{s}_2\|_2 < D_{d+1}) \\ \wedge (A_a < \angle(\mathbf{s}_1, \mathbf{s}_2) < A_{a+1}) \}$$

with the new dimension a running from 1 to N_a (number of angle bins).

The size of the matrix is $N_c \times N_c \times N_d \times N_a$. For 20 clusters, 10 distance bins and 4 angle bins, this leads to a feature vector of size 16000. The classification is done with the help of a multi-class SVM, running in a **one-vs-rest** mode, and using a histogram-intersection kernel. The libSVM¹ implementation¹ was used for the experiments. Despite the large dimensionality, the speed should be acceptable for most cases. Training the SVM classifier with 9000 examples takes about 2 hours of computation time, the final classification of 1000 examples just under 10 minutes.

3.3 Results and Conclusion

We submitted two runs for the ImageCLEF medical annotation task. Both runs used the orientation variant accumulator defined in Section 3.2 and used vectors of the same size, with $N_c = 20$, $N_d = 10$ and $N_a = 4$. The only difference was in the number of salient points, N_s , selected per image. The first run used $N_s = 1000$, while the second used $N_s = 800$. For the given test set, the runs achieved an error rate of 16.7 % and 17.9 % respectively, which means that our best run was ranked second overall for the task. There might have been some room left for better parameter tuning, but it was not performed due to lack of resources, and because there are quite a few parameters to tune. Thus, instead of a joint optimization, each parameter was only individually selected through empirical methods. The presence of a large number of tunable parameters could be considered as a drawback of the proposed approach. However, this also means that, given sufficient computation power, the system could automatically learn to classify many different configurations of classes through cross-validation.

Figure 3 shows the nearest neighbour results (L_1 norm) for four sample query images. It can be seen that in most cases, the features are very suited for the task. Local features are in general robust against occlusion and partial matching, which is in general advantageous for a recognition task. This is also true for the radiograph annotation task described in this paper, but there are exceptions. For example, the top left box contains a chest radiograph query, and the radiographs with a complete frontal chest and a partial frontal chest

¹ <http://lmb.informatik.uni-freiburg.de/lmbsoft/libsvm/>

Table 1. Confusion Matrix for the four categories as shown in Section 2

Class	108	109	110	111	Accuracy
108	78	2	0	12	84.7 %
109	0	4	0	0	100 %
110	0	0	6	1	85.7 %
111	10	0	0	184	94.8 %

apparently belong to different classes (class number 108 and 29 respectively). In other words, robustness or invariance towards a particular transformation might not always be good for the classification accuracy. In such cases, it can be beneficial to use region specific extra features, once it is determined that the radiograph is of a particular type e.g. one of a chest. An automatic feature selection method could also be used to select discriminative features from the large set of features.

A possibility that could be tried is to generate so called virtual samples by subjecting the training vectors to transformations that are expected in reality. This is especially useful if the original training database is not expressive enough for all possible variations. Generating relational features for slightly rotated images is straightforward and does not require recomputation of the features. It is sufficient to circularly shift the relational feature vector obtained for each (r_1, r_2, ϕ) combination to achieve the desired effect.

In the end, we note another interesting observation, namely in the confusion matrix for classes which appear very similar to an untrained human being, for example, between the classes numbered 108 through 111, as shown in Figure 2. Surprisingly the system is able to distinguish these quite effectively. A possible explanation could be that the classes differ in their local structure which is not visible to the naked eye.

References

1. Guelld, M.O., Kohnen, M., Keyzers, D., Schubert, H., Wein, B.B., Bredno, J., Lehmann, T.M.: Quality of DICOM header information for image categorization. In: Proc. SPIE, Medical Imaging 2002, vol. 4685, pp. 280–287 (2002)
2. Schulz-Mirbach, H.: Invariant features for gray scale images. In: Sagerer, G., Posch, S., Kummert, F. (eds.) 17. DAGM - Symposium “Mustererkennung”, Reihe Informatik aktuell, Bielefeld, pp. 1–14. Springer, Heidelberg (1995)
3. Louprias, E., Sebe, N.: Wavelet-based salient points for image retrieval (1999)
4. Schael, M.: Methoden zur Konstruktion invarianter Merkmale für die Texturanalyse. PhD thesis, Albert-Ludwigs-Universität, Freiburg (June 2005)
5. Ojala, T., Pietikäinen, M., Mäenpää, T.: Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 404–420. Springer, Heidelberg (2000)

6. Setia, L., Teynor, A., Halawani, A., Burkhardt, H.: Image Classification using Cluster-Cooccurrence Matrices of Local Relational Features. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR 2006), Santa Barbara, CA, USA (October 26-27, 2006)
7. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Mller, H.: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain

MorphoSaurus in ImageCLEF 2006: The Effect of Subwords On Biomedical IR

Philipp Daumke, Jan Paetzold, and Kornel Marko

University Hospital Freiburg, Dept. of Medical Informatics,
Stefan-Meier-Str. 26, 79106 Freiburg, Germany
philipp.daumke@klinikum.uni-freiburg.de,
jan.paetzold@uniklinik-freiburg.de, kornel.marko@uni-freiburg.de
<http://www.morphosaurus.net>

Abstract. In the 2006 ImageCLEF Medical Image Retrieval task we evaluate the effects of deep morphological analysis for mono- and cross-lingual document retrieval in the biomedical domain. The morphological analysis is based on the MORPHOSAURUS system in which subwords are introduced as morphologically meaningful word units. Subwords are organized in language specific lexica that were partly manually and partly automatically generated and currently cover six European languages. They are linked together in a multilingual thesaurus. The use of subwords instead of full words significantly reduces the number of lexical entries that are needed to sufficiently cover a specific language and domain. A further benefit of the approach is its independence from the underlying retrieval system. We combined MORPHOSAURUS with the open-source search engine *Lucene* and achieved precision gains of up to 25% over the baseline for a monolingual setting and promising results in a multilingual scenario.

Keywords: Cross Language Information Retrieval, Biomedical Information Retrieval, Subword Indexing, Morpho-Semantic Indexing, Stemming.

1 Introduction

The conventional view on human language is word-centered, at least for written language in which words are clearly delimited by spaces. It builds on the hypothesis that words are the basic building blocks of phrases and sentences. In syntactic theories, words constitute the terminal symbols. Therefore, it appears straightforward to break down natural language to the word level. However, looking at the sense of natural language expressions, evidence can be found that semantic atomicity frequently does not coincide with the word level, which bears methodical challenges even for pretended ‘simple’ tasks such as tokenization of natural language input [1]. As an example, considering the English noun phrase “*high blood pressure*”, the word limits reflect quite well the semantic composition, whereas this is not the case in its literal translations “*verhoogde bloeddruk*” (Dutch), “*högt blodtryck*” (Swedish) or “*Bluthochdruck*” (German). Especially in

technical sublanguages such as the medical one, atomic senses are encountered at different levels of fragmentation or granularity. An atomic sense may correspond to word stems (e.g., “*hepat*” referring to “*liver*”), prefixes (e.g., “*anti-*”, “*hyper-*”), suffixes (e.g., “*-logy*”, “*-itis*”), larger word fragments (“*hypophys*”), words (“*spleen*”, “*liver*”) or even multi-word terms (“*yellow fever*”). The possible combinations of these word-forming elements are immense and ad-hoc term formation is common.

Extracting atomic sense units from texts as a basis for the semantic interpretation of natural language is therefore an important goal in various areas of research dealing with natural language processing, such as information retrieval, text mining or speech recognition. Especially in morphologically rich languages (German, Finnish, Swedish, etc.), where classical rule-based stemmers such as the Porter Stemmer[2] are not particular effective, a deeper morphological analysis is widely acknowledged to improve the retrieval performance in IR systems [3,4,5]. However, the lack of accuracy of current unsupervised approaches for morphological analysis often hampers the automatic or semi-automatic creation of an underlying knowledge base. We developed such a lexical resource and introduced the notion of subwords [6], i.e. self-contained, semantically minimal units. Language-specific subwords are linked by intralingual as well as interlingual synonymy and grouped together in terms of concept-like language independent equivalence classes. For ImageCLEFmed 2006, we prove the positive effect of using subwords on the monolingual and multilingual biomedical text retrieval tasks.

2 The MorphoSaurus-System

2.1 Subword Lexicons and Subword Thesaurus

Subwords are assembled in a multilingual dictionary and thesaurus, which contain their entries, special attributes and semantic relations between them, according to the following considerations:

- Subwords are listed, together with their attributes such as language (English, German, Portuguese) and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned one or more morpho-semantic identifier(s) representing the corresponding synonymy class, the MORPHOSAURUS identifier (MID). Intra- and interlingual semantic equivalence are judged within the context of medicine only.
- Semantic links between synonymy classes are added. We subscribe to a shallow approach in which semantic relations are restricted to a single paradigmatic relation has-meaning, which relates one ambiguous class to its specific readings¹, and a syntagmatic relation expands-to, which consists of predefined segmentations in case of utterly short subwords².

¹ For instance, {*head*} $\Rightarrow$ {*zephal,kopf,caput,cephal,cabec,cefal*} OR {*leader,boss,lider,cheffe*}

² For instance, {*myalg*} $\Rightarrow$ {*muscle,muskel,muscul*} $\oplus$ {*schmerz,pain,dor*}

Compared to relationally richer, e.g., WORDNET-based, interlinguas as applied for cross-language information retrieval [7,8], our set of semantic relations is obviously limited. We refrain from introducing hierarchical relations between MIDs because such links can be acquired from domain-specific vocabularies, e.g. the Medical Subject Headings (MeSH)[9], cf. also [10] for the mapping of MIDs to appropriate MeSH terms).

The manual construction of our lexical resources was done over the last five years with a changing amount of man power. The English, German and Portuguese subword lexicons were created fully manually, while for French, Spanish and Swedish, additional machine learning techniques were applied in order to bootstrap the lexical resources for these languages[11]. Overall, the lexicons contain 97,249 entries³, corresponding to 21,679 equivalence classes. In terms of relations between equivalence classes, there are currently 497 distinct *expands-To* and 1,369 distinct *has-Sense* relations defined in the thesaurus.

2.2 Morpho-Semantic Normalization

Figure 1 depicts how source documents (top-left) are converted into an interlingual representation by a three-step morpho-semantic indexing procedure. First, each input word is orthographically normalized in terms of lower case characters and according to language-specific rules for the transcription of diacritics (top-right). Next, words are segmented into sequences of subwords or left as is when no subwords can be decomposed (bottom-right). The segmentation results are checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g., segmentations without stems or ones beginning with a suffix). Finally, each meaning-bearing subword is replaced by a language-independent semantic identifier, its MID, thus producing the interlingual output representation of the system (bottom-left). In Figure 1, MIDs which co-occur in both document fragments appear in bold face. A comparison of the original natural language document input (top-left) and their interlingual representation (bottom-left) already reveals the degree of (hidden) similarity uncovered by the overlapping MIDs.

3 Experiments

3.1 Indexing and Query Preparation Process

In this year's ImageCLEFmed we combined the MORPHOSAURUS system with the open-source search engine Lucene,⁴ a high performance, scalable, cross-platform retrieval system. In the preparation phase all annotations of the

³ For comparison, the size of WORDNET [12] assembling the *lexemes* of general English in the 2.0 version is on the order of 152,000 entries (<http://www.cogsci.princeton.edu/~wn/>). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

⁴ <http://jakarta.apache.org/lucene/docs/index.html>

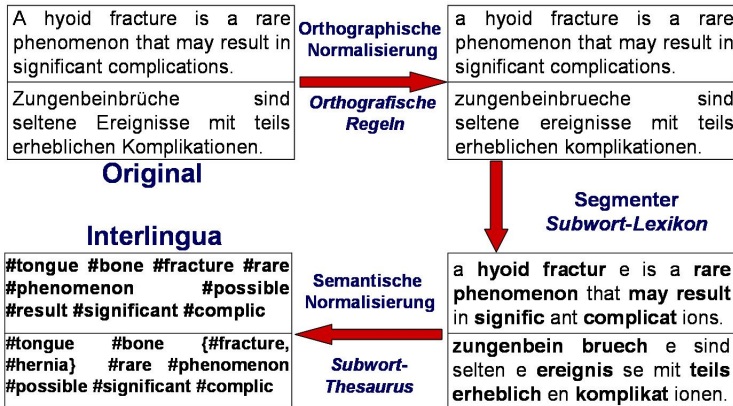


Fig. 1. Morpho-Semantic Indexing Pipeline

whole dataset containing textual information from Casimage, MIR, PEIR, and PathoPIC datasets were extracted. For all image annotations, we created four Lucene fields which could be queried separately. The first field contained headlines, keywords and additional concise XML tags in the original representation (*orig.t*). The second field contained all other free text information that was extracted from the dataset (*orig.d*). The third and fourth field contained the MORPHOSAURUS representation of the corresponding first and second field (*mid.t* and *mid.d*). In order to compare the subword approach with traditional automated stemming routines we added two further fields (*stem.t* and *stem.d*) containing the annotations processed by the Porter Stemmer⁵. In addition to these six query fields, a language flag was added for each image annotation (*language*).

In the topics collection, "Show me", "Zeige mir" and "Montre-moi des" and language specific stopwords⁶ were removed. The queries were subsequently transformed into the MORPHOSAURUS interlingua resulting in 30 "*orig/stem/mid*" query triplets, each represented in all three languages.

3.2 Monolingual Scenario / Textual

One of our goals was to show the effectiveness of our subword approach for monolingual biomedical text retrieval. We therefore created a scenario which makes use of the English subset only. Four different runs were prepared:

1. *Orig-En-En*: As a baseline we took the original english query and searched in the corresponding two Lucene fields which contained the original English annotations (*orig.t*, *orig.d*, *language : en*).

⁵ We used the stemmer available on <http://www.snowball.tartarus.org> (last visited on Dec. 2006).

⁶ The Snowball Stemmer incorporates stop word lists containing 172 English, 232 German and 155 French entries.

2. *Stem-En-En*: These queries were created by processing the original queries with the Porter Stemmer and sent to the corresponding (*stem_t*, *stem_d*, *language : en*)-fields of Lucene.
3. *Mids-En-En*: In this run we took the morpho-semantic normalized form of each English query and searched in the two fields that contained the MID representation of the document annotations (*mid_t*, *mid_d*, *language : en*).
4. *Both-En-En*: Here, we combined both the original query and the morpho-semantic normalized query in a simple disjunct fashion. The original words were queried in the original fields (*orig_t* and *orig_d*, *language : en*) and the normalized queries in the MID fields (*mid_t*, *mid_d*, *language : en*).

3.3 Multilingual Scenario / Textual and Mixed

In the multilingual scenario, only the morpho-semantic representation of the topics in English, German and French were used to search in both MID fields (*mid_t* and *mid_d*). No restrictions to the language flag of the document collections were made. The resulting runs were named *Mids-En-All*, *Mids-De-All* and *Mids-Fr-All*, respectively. As a baseline, the original queries in all languages were used to separately search in the language corresponding original document fields (i.e. English queries in English documents, German queries in German documents and French queries in French documents). The baseline is referred to as *Orig-All-All*. As a second baseline, we also consider the monolingual baseline *Orig-En-En* as a solid reference value, taking into account that the English subset represents a comprehensive part of the annotations (almost 93% of all annotations are available in English).

In the mixed multilingual scenario we combined the retrieval results of the multilingual textual scenario with the GIFT visual retrieval results. In addition, a test run with the best textual run *Both-En-En* in combination with the GIFT results was carried out, as we expected this run to be the best of all runs. As the textual retrieval results were expected to perform better than the visual retrieval results, we ranked those documents at top which were found in both retrieval systems under the first 50 hits, followed by only textual retrieval results. Hits only found in the visual retrieval system were discarded. The runs are named *Orig-All-All-Comb*, *Both-En-En-Comb*, *Mids-En-All-Comb*, *Mids-De-All-Comb* and *Mids-Fr-All-Comb*.

4 Results

Interesting from a realistic retrieval perspective, at least to our view, is the average gain on the top two recall points as well as the P5 and P20 values. Together with the overall *map* value, these figures can be found in Table 1 for the monolingual and in Table 2 for the multilingual scenario.

The first observation in the **monolingual scenario** was that the simple stemming approach (*Stem-En-En*) does not perform better than the baseline *Orig-En-En*, with values between 91% (*map*) and 111% (*P20*) of the baseline. Also,

Mids-En-En performs not as good as the baseline *Orig-En-En* regarding *map* (0.13 vs.016) and *top2avg* (0.32 vs. 0.38) and achieves same results as *Orig-En-En* for the P5 (0.49) and P20 (0.36) precision values.

The *Both-En-En* run outperforms all other runs in all relevant values. Regarding the P5 and P20 values of both *Both-En-En* and the baseline run *Orig-En-En* (P5: 0.60 vs. 0.48, P20: 0.45 vs. 0.36), *Both-En-En* exceeds the baseline by up to 25%. *Top2avg* (0.45 vs. 0.38) is 20% higher and the *map*-value (0.18 vs. 0.16) is still 10% increased. The P5 value of *Both-En-En* is second best of all (automatic/textual) runs at ImageCLEFmed 2006.

Table 1. Standard Precision/Recall Table for the Monolingual Scenario

MonoText							
Scenario	Orig-En-En	Stem-En-En	Mids-En-En	Both-En-En			
map	0.1625	0.1482	91%	0.1297	80%	0.1792	110%
top2avg	0.3778	0.3669	97%	0.3175	84%	0.4525	120%
P5	0.4867	0.4667	96%	0.4867	100%	0.5933	122%
P20	0.3600	0.4000	111%	0.3550	99%	0.4500	125%

Table 2. Standard Precision/Recall Table for the Multilingual Scenario (Textual)

Multilingual / Textual									
Scenario	Orig-All-All	Orig-En-En	Mids-En-All	Mids-De-All	Mids-Fr-All				
map	0.1068	0.1625	152%	0.1366	128%	0.1439	135%	0.0734	69%
top2avg	0.2881	0.3778	131%	0.3970	138%	0.3751	130%	0.2028	70%
P5	0.3200	0.4867	152%	0.4200	131%	0.4000	125%	0.3103	97%
P20	0.2600	0.3600	138%	0.3467	133%	0.3383	130%	0.2293	88%

Regarding the **multilingual runs**, the first observation was that the multilingual baseline *Orig-All-All* was notably lower than the monolingual baseline (the monolingual results of *Orig-En-En* are added in Table 2 for comparison). Obviously, due to the comprehensive part of English annotations (93% of the annotations are available in English) querying the English subset achieves better results than querying the German and French subset. Regarding the multilingual runs, we observe similar results for the German and the English test scenarios with P5 values between 0.40 (*Mids-De-All*) and 0.42 (*Mids-En-All*) and *top2avg* values between 0.38 (*Mids-De-All*) and 0.40 (*Mids-En-All*). Both runs exceeded the multilingual baseline regarding all relevant values by about 25% to 38%.

Considering the low number of German and French annotations compared to English, the *Mids-De-All* run can nearly be considered a fully *translingual*

Table 3. Standard Precision/Recall Table for the Multilingual Scenario (Mixed). *Comb*-suffixes of the scenario identifiers were stripped for lack of space

Multilingual / Mixed									
Scenario	Orig-All-All	Both-En-En	Mids-En-All	Mids-De-All	Mids-Fr-All				
map	0.1079	0.1791	166%	0.1383	128%	0.1441	134%	0.0746	69%
top2avg	0.2942	0.4516	154%	0.4015	136%	0.3726	127%	0.2020	69%
P5	0.3333	0.6000	180%	0.4400	132%	0.4000	120%	0.3241	97%
P20	0.2633	0.4500	171%	0.3517	134%	0.3350	127%	0.2259	86%

run (German queries on English documents). It is particular promising that this run performed as good as the English run (*Mids-En-All*) and the monolingual baseline (*Orig-En-En*) and clearly better than the multilingual baseline (*Orig-All-All*). The French run *Mids-Fr-All* performed relatively poor compared to the others. This reflects the fact that we only just started to build the French Subword lexicon.

The **mixed multilingual runs** (Table 3) performed in average not better than the multilingual textual runs. Obviously, better merging algorithms to better exploit the synergies between the textual and the visual runs are needed and are due in future work.

While the P5, P20 and top2avg values of our best runs are quite high compared with other groups([13]), the overall map value is mean. This is due to an overbalance of certain unspecific MIDs in our subword approach that cause a decrease of precision values beginning roughly at the cut-off point 100. However, as the P100-P1000 values are of only limited value for a user in real life scenario, these findings are not too worrying. Still the increase of the overall map value is an important goal of our future work.

5 Conclusion and Future Work

We introduced a subword-based approach for biomedical text retrieval that addresses many of the general and domain specific challenges in the current CLIR research. In our monolingual test runs, we showed that a combination of original and morpho-semantic normalized queries remarkably boosts precision up to 25% (P5, P20), compared to the baseline. The best of our runs was second best of all (automatic/textual) test runs regarding the P5 value. In the multilingual runs we achieve similar results for the English and German test runs. In all scenarios, the P5, P20 and top2avg values are distinctly higher than the multilingual baseline. A detailed analysis of the test runs is now due in order to determine which part of our system contributed to which extend to the precision gain. Also, an increase of the overall *map*-value is an important goal in our future work. A medium-term goal is to show the usefulness of subwords in other (technical) domains.

References

1. Grefenstette, G., Tapanainen, P.: What is a word, what is a sentence? problems of tokenization. In: *Proceedings of the 3rd International Conference on Computational Lexicography*, pp. 79–87 (1994)
2. Porter, M.F.: An algorithm for suffix stripping. *Program* 14(3), 130–137 (1980)
3. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: an International Journal* 41(3), 523–547 (2005)
4. Pirkola, A.: Morphological typology of languages for IR. *Journal of Documentation* 57(3), 330–348 (2001)
5. Braschler, M., Ripplinger, B.: How effective is stemming and compounding for German text retrieval? *Information Retrieval* 7(3-4), 291–316 (2004)
6. Markó, K., Schulz, S., Hahn, U.: MorphoSaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine* 44(4), 537–545 (2005)
7. Gonzalo, J., Verdejo, F., Chugur, I.: Using EuroWordNet in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence* 13(7), 647–678 (1999)
8. Ruiz, M., Diekema, A., Sheridan, P.: Cindor conceptual interlingua document retrieval: Trec-8 evaluation In: Voorhees, E.M., Harman, D.K. (eds.) *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, Gaithersburg, MD, November 17-19, 1999, pp. 500–246 (National Institute of Standards and Technology (NIST) (1999) 597-606 NIST Special Publication)
9. MESH: Medical Subject Headings. National Library of Medicine, Bethesda, MD (2005)
10. Markó, K., Daumke, P., Schulz, S., Hahn, U.: Cross-language MeSH indexing using morpho-semantic normalization. In: Musen, M.A. (ed.) *AMIA'03 - Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications*, Washington, D.C, November 8-12, 2003, pp. 425–429. Hanley & Belfus, Philadelphia, PA (2003)
11. Markó, K., Schulz, S., Medelyan, A., Hahn, U.: Bootstrapping dictionaries for cross-language information retrieval. In: *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, August 15-19, 2005, pp. 528–535. ACM, New York (2005)
12. Fellbaum, C.: *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge (1998)
13. Mueller, H., Deselaers, T., Lehmann, T., Clough, P., Hersch, W.: Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In: *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain. LNCS (to appear, 2006)

Medical Image Retrieval and Automated Annotation: OHSU at ImageCLEF 2006

William Hersh, Jayashree Kalpathy-Cramer, and Jeffery Jensen

Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University
Portland, OR, USA
hersh@ohsu.edu

Abstract. Oregon Health & Science University participated in both the medical retrieval and medical annotation tasks of ImageCLEF 2006. Our efforts in the retrieval task focused on manual modification of query statements and fusion of results from textual and visual retrieval techniques. Our results showed that manual modification of queries does improve retrieval performance, while data fusion of textual and visual techniques improves precision but lowers recall. However, since image retrieval may be a precision-oriented task, these data fusion techniques could be of value for many users. In the annotation task, we assessed a variety of learning techniques and obtained classification accuracy of up to 74% with test data.

1 Introduction

Oregon Health & Science University (OHSU) participated in the medical image retrieval task and automatic annotation tasks of ImageCLEF 2006 (Müller, Deselaers et al., 2006). Similar to our approach for the medical retrieval tasks in 2005, we focused on two primary areas: the manual modification of queries and the use of visual information in series with the results of the textual searches.

The automatic annotation task provided a forum to evaluate our algorithms for medical image classification and annotation. We evaluated the performance of a variety of low level image features using a two-layer neural network architecture for the classifier. This classifier was then used in conjunction with textual results of the image retrieval system to improve the precision of the search for the medical retrieval tasks.

2 Image Retrieval

The goal of the ImageCLEF medical image retrieval task is to retrieve relevant images from a test collection of about 50,000 images that are annotated in a variety of formats and languages (Müller, Deselaers et al., 2006). Thirty topics were developed, evenly divided as amenable to textual, visual, or mixed retrieval techniques. The top-ranking images from runs by all participating groups were judged as definitely, possibly, or not relevant by relevance judges.

The mission of information retrieval research at Oregon Health & Science University (OHSU) is to better understand the needs and optimal implementation of systems for users in biomedical tasks, including research, education, and clinical care. The goals of the OHSU experiments in the medical image retrieval task of ImageCLEF were to assess manual modification of topics with and without visual retrieval techniques. We manually modified the topics to generate queries, and then used what we thought would be the best run (which in retrospect was not) for combination with visual techniques, similar to the approach we took in ImageCLEF 2005 (Jensen and Hersh, 2005).

2.1 System Description

Our retrieval system was based on the open-source search engine, Lucene (Gospodnetic and Hatcher, 2005), which is part of the Apache Jakarta distribution. We have used Lucene in other retrieval evaluation forums, such as the Text Retrieval Conference (TREC) Genomics Track (Cohen, Bhupitiraju et al., 2004; Cohen, Yang et al., 2005). Documents in Lucene are indexed by parsing of individual words and weighting of those words with an algorithm that sums for each query term in each document the product of the term frequency (TF), the inverse document frequency (IDF), the boost factor of the term, the normalization of the document, the fraction of query terms in the document, and the normalization of the weight of the query terms, for each term in the query. The score of document d for query q consisting of terms t is calculated as follows:

$$score(q, d) = \sum_{t \in q} tf(t, d) * idf(t) * boost(t, d) * norm(d, t) * frac(t, d) * norm(q)$$

where:

- $tf(t, d)$ = term frequency of term t in document d
- $idf(t)$ = inverse document frequency of term t
- $boost(t, d)$ = boost for term t in document d
- $norm(t, d)$ = normalization of d with respect to t
- $frac(t, d)$ = fraction of t contained in d
- $norm(q)$ = normalization of query q

As Lucene is a code library and set of routines for IR functionality, it does not have a standard user interface. We have therefore also created a search interface for Lucene that is tailored to the ImageCLEF medical retrieval test collection structure (Hersh, Müller et al., 2006) and the ability to use the MedGIFT search engine for visual retrieval on single images (Müller, Geissbühler et al., 2005). We did not use the user interface for these experiments, though we plan to undertake interactive user experiments in the future.

2.2 Runs Submitted

We submitted three general categories of runs:

- Automatic textual - submitting the topics as phrased in the official topics file directly into Lucene. We submitted each of the three languages in separate runs,

along with a run that combined all three languages into a single query string and another run that included the output from the Babelfish translator¹

- Manual textual - manually editing of the official topic files by one of the authors (WRH). The editing mostly consisted of removing function and other common words. Similar to the automatic runs, we constructed query files in each of the three languages, along with a run that combined all three languages into a single query string and a final run that included the output from the Babelfish translator.
- Interactive mixed - a combination of textual and visual techniques, described in greater detail below.

The mixed textual and visual run was implemented as a serial process, where the results of what we thought would be our best textual run were passed through a set of visual retrieval steps. This run started by using the top 2000 retrieved images of the OHSU_all textual run. These results were combined with the top 1000 results distributed from the medGIFT (visual) system. Only those images that were in both lists were chosen. These were ordered by the textual ranking, with typically 8 to 300 images in common.

A neural network-based scheme using a variety of low level, global image features was used to create the visual part of the retrieval system. The retrieval system was created in MATLAB² using Netlab³ (Bishop, 1995; Nabney, 2004). We used a multi-layer perceptron architecture to create the two-class classifiers to determine if a color image was a 'microscopic' image or 'gross pathology.' It was a two layer structure, with a hidden layer of approximately 50-150 nodes. A variety of combinations of the image features were used as inputs. All inputs to the neural network (the image feature vectors) were normalized using the training set to have a mean of zero and variance of 1.

Our visual system then analyzed the sample images associated with each sub-task. If the query image was deemed to be a color image by the system, the set of top 2000 textual images was processed and those that were deemed to be color were moved to the top of the list. Within that, the ranking was based on the ranking of the textual results.

A neural network was created to process color images to determine if they were microscopic or gross pathology/photograph. The top 2000 textual results were processed through this network and the appropriate type of image (based on the query image) received a higher score. Relevance feedback was used to improve the training for the network (Crestani, 1994; Han and Huang, 2005; Wang and Ma, 2005). Low level texture features based on grey-level co-occurrence matrices (GLCM) were used as input to the neural network (Haralick, 1979; Rahman, Desai et al., 2005). We also created neural networks for a few classes of radiographic images, based on the system that we had used for the automatic annotation class (described in detail in the next section). Images identified as being of the correct class received a higher score.

The primary goal of these visual techniques was to move the relevant images higher on the ordered list of retrieved images, thus leading to higher precision.

¹ <http://babelfish.altavista.com>

² <http://www.mathworks.com>

³ <http://www.ncrg.aston.ac.uk/netlab/index.php>

However, we would be limited in the recall to only those images that had already been retrieved by the textual search. Thus, even in the ideal case, where all the relevant images were moved to the top of the list, the MAP would be limited by the number of relevant images that were retrieved by the textual search (recall of the textual search).

We improved our modality classifier after submitting our official runs to ImageCLEF. We can now classify images into 6 categories for color images (stained histology (microscopic) images, photographs and gross pathology images, electroencephalographical images (EEGs) and electrocardiograms (ECGs), Powerpoint slides, as well as a few miscellaneous images. Grey-scale images are also classified into modalities including angiography, computerized tomography scans (CT), X-ray, Magnetic resonance (MRI), ultrasound, and scintigraphy. These classifiers, achieving >95% accuracy, were tested on a random subset of the ImageCLEFmed topics.

2.3 Results and Analysis

The characteristics of the submitted OHSU runs are listed in Table 1, with various results shown in Figure 1. The automatic textual runs were our lowest scoring runs. The best of these runs was the English-only run passed through the Babelfish translator, which obtained a MAP of 0.1264. The remaining runs all performed poorly, with all MAP results under 0.08. The manual textual runs performed somewhat better. Somewhat surprising to us, the best of these runs was the English-only run (OHSUeng). This was our best run of all, with a MAP of 0.2132. It outperformed an English-only run with terms from automatic translation added (OHSUeng_trans, with a MAP of 0.1906) as well as a run with queries of topic statements from all languages (OHSU_all, with a MAP of 0.1673).

The MAP for our interactive-mixed run, OHSU_m1, was 0.1563. As noted above, this run was based on modification of OHSUall, which had a MAP of 0.1673. At a first glance, it appears that performance was worsened with the addition of visual techniques, due to the lower MAP. However, as seen in Figure 1, and similar to our results from 2005, the average precision at various numbers of images retrieved was higher, especially at the top of the retrieval list. This confirmed our finding from 2005 that visual techniques used to modify textual runs diminish recall-oriented measures like MAP but improve precision at the very top of output list, which may be useful to real users. There was a considerable variation in performance on different topics. For most topics, the addition of visual techniques improved early precision, but for some, the reverse was true.

We also looked at MAP for the tasks separated by their perceived nature of the question (one favoring visual, semantic, or mixed techniques, see Table 2). For the visual and mixed queries, the incorporation of visual techniques improved MAP. However, for semantic queries, there was a serious degradation in MAP by the addition of the visual steps in the retrieval process. This, however, is driven by only one query, number 27, where MAP for OHSU_all was 0.955, while for OHSU_m1 was 0.024. Excluding this query, MAP for OHSU_m1 was 0.161 while that of OHSU_all was 0.140, indicating a slight improvement as a result of the addition of visual techniques.

Table 1. Characteristics of OHSU runs

Run_ID	Type	Description
OHSU_baseline_trans	Auto-Text	Baseline queries in English translated automatically
OHSU_english	Auto-Text	Baseline queries in English only
OHSU_baseline_notrans	Auto-Text	Baseline queries in all languages
OHSU_german	Auto-Text	Baseline queries in German only
OHSU_french	Auto-Text	Baseline queries in French only
OHSUeng	Manual-Text	Manually modified queries in English
OHSUeng_trans	Manual-Text	Manually modified queries in English translated automatically
OHSU-OHSUall	Manual-Text	Manually modified queries in all three languages
OHSUall	Manual-Text	Manually modified queries in all three languages
OHSUger	Manual-Text	Manually modified queries in German
OHSUfre	Manual-Text	Manually modified queries in French
OHSU-OHSU_m1	Interactive-Mixed	Manually modified queries filtered with visual methods

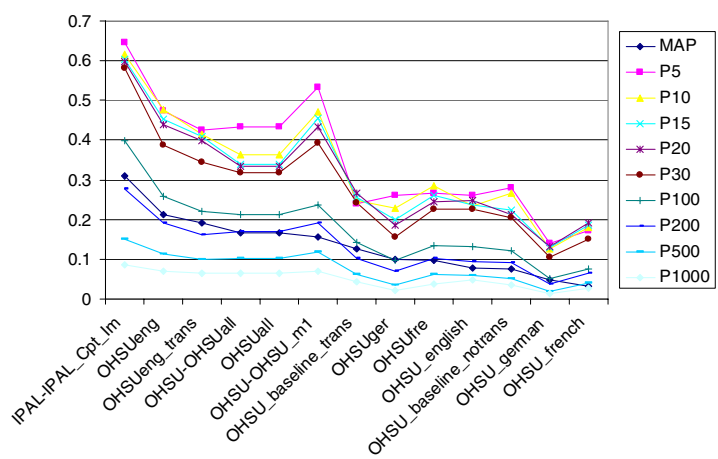


Fig. 1. MAP and precision at various retrieval levels for all OHSU runs and the run with the best overall MAP from ImageCLEFmed 2006, IPAL-IPAL_Cpt_Im

Table 2. MAP by query type for mixed and textual queries

Query Type	MAP	
	OHSU_m1	OHSUall
Visual	0.139	0.128
Mixed	0.182	0.148
Semantic	0.149	0.226

Table 3. Modality of images returned using textual query for topic 25, *Show me microscopic images of tissue from the cerebellum*

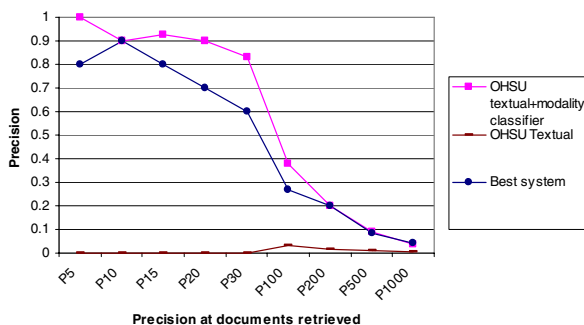
Image type	Number of images
Total returned by textual query	2000
Grey-scale	1476
Photograph/gross pathology	408
Microscope	116

The new color modality classifier was tested on a small random subset of the ImageCLEFmed 2006 topics. Analysis of our textual results indicated that in many queries, especially those of a visual or mixed nature, up to 75% of the top 1000 results were not of the correct modality. An example of this is shown in Table 3. Only 90 of the top 2000 images returned by the textual query alone were of the desired modality. The precision of this search was improved by the use of the modality detector as seen in Figure 2.

Our runs demonstrated that manual modification of topic statements makes a large performance difference, although our results are not as good as some groups that did automatic processing of the text of topics. Our results also showed that visual retrieval techniques provide benefit at the top of the retrieval output, as demonstrated by higher precision at various output levels, but are detrimental to recall, as shown by lower MAP.

3 Automated Image Annotation

The goal of this task was to correctly classify 1000 radiographic medical images into 116 categories (Müller, Deselaers et al., 2006). The images differed in the “modality, body orientation, body region, and biological system examined,” according to the track Web site. The task organizers provided a set of 9,000 *training* images that were classified into these 116 classes. In addition, another set of classified images (numbering 1000) was provided as a *development* set. The development set could be used to evaluate the effectiveness of the classifier prior to the final evaluation.

**Fig.2.** Improvement of precision for topic 25 using the image modality classifier in series with textual results

We used a combination of low-level image features and a neural network based classifier for the automated image annotation task. Our results were in the middle of the range of results obtained for all groups, indicating to us the potential capabilities of these techniques as well as some areas of improvement for further experiments.

3.1 System Description

A neural network-based scheme using a variety of low-level, largely global, texture and histogram image features was used to create the classifier. The system was implemented in MATLAB using the Netlab toolbox. All images were padded to create a square image and then resized to 256x256 pixels. A variety of features described below were tested on the development set. These features were combined in different ways to try to improve the classification ability of the system, with the final submissions were based on the three best combinations of image features. The features included:

- *Icon*: A 16x16 pixel 'icon' of the image was created by resizing the image using bilinear extrapolation.
- *GLCM*: Four gray level co-occurrence matrices (GLCM) (Haralick, 1979) matrices with offsets of 1 pixel, 0, 45, 90 and 135 degrees were created for the image after rescaling the image to 16 levels. GLCM statistics of contrast, correlation, energy, homogeneity and entropy were calculated for each matrix. A 20 dimensional vector was created for each image by concatenating the 5 dimensional vector obtained by each of the four offset matrices.
- *GLCM2*: In order to capture the spatial variation of the images in a coarse manner, the resized image (256x256) was partitioned into 5 squares of size 128x128 pixels (top left, top right, bottom left, bottom right, centre). A gray level correlation matrix was created for each partition. The five 20-dimensional vectors from each of the partitions were concatenated to create a feature vector of dimension 100.
- *Hist*: A 32-bin histogram was created for each image and counts were used as the input

We used a multilayer perceptron architecture to create the multi-class classifier (Bishop, 1995; Nabney, 2004). It was a two layer structure, with a hidden layer of approximately 200-400 nodes. A variety of combinations of the above image features were used as inputs. All inputs to the neural network (the image feature vectors) were normalized using the training set to have a mean of zero and variance of 1. The architecture was optimized using the training and development sets provided.

The network architecture, primarily the number of hidden nodes, needed to be optimized for each set of input feature vectors, since the length of the feature vectors varied from 32 to 356. The training set was used to create the classifier, typically with the accuracy increasing with an increase in the number of hidden nodes. It was relatively easy to achieve 100% classification accuracy on the training set. However, there were issues with overfitting if too many hidden nodes were used. We used empirical methods to optimize the network for each set of feature vectors by using a

network architecture that resulted in the highest classification accuracy for the development set.

3.2 Runs Submitted

We submitted four runs, iconGLCM2 using just the training set for creating the net, iconGLCM2 using the development and training set for creating the net, iconHist, and iconHistGLCM.

3.3 Results and Analysis

The best results for the development set were obtained using a 356 dimensional normalized input vector consisting of the icon (16x16) concatenated with the GLCM vectors of the partitioned image. The classification rate on the training set was 80%. The next best result was obtained using a 288 dimensional normalized input vector consisting of the icon (16x16) concatenated with Hist. The classification rate on the development set was 78%. Most other runs including just the icon or GLCM2 gave about 70-75% classification accuracy, as seen in Table 4. However, the results obtained on the test set were lower than those of the development set.

A few classes were primarily responsible for the differences seen between the development set and test set. Class 108 saw the most significant difference. Most of the misclassification of class 108 was into class 111, visually a very similar class. Observing the confusion matrices in general for all the runs, the most misclassifications occurred between classes 108 and 111, and 2 and 56.

Table 4. Classification rates for OHSU automatic annotation runs

Feature vector	Classification rate	
	Development	Test
iconHist	78	69
iconGLCMHist	78	72
iconGLCM2	80	74

Following our availability of the results, we performed additional experiments to improve the classification between these sets of visually similar classes. We created two new classifiers to distinguish between class 2 and 56, and between class 108 and 111. We merged images labeled by the original classifier as class 2 and 56, and class 108 and 111 and then applied the new classifiers on these newly merged classes. Using this hierarchical classification, we improved our classification accuracy by about 4% (to 79%) overall for the test set. This seems like a promising approach to improve the classification ability of our system.

One of the issues with the database is that the number of training images in each of the classes is quite varied. Another issue is that there are some classes that are visually quite similar while other classes that have quite a bit of within class variation. These issues were proved to be a little challenging for our system.

4 Conclusions

Manual modification of topic statements improved our performance for the image retrieval task. Inclusion of visual techniques in series with our textual result increased precision, but was detrimental to recall, depending on the techniques used. However, for most image retrieval tasks, precision may be more important than recall, so visual techniques may be of value in real-world image retrieval systems. Additional research on how real users query image retrieval systems could shed light on which system-oriented evaluation measures are most important.

Also suggested by our runs is that system performance is dependent upon the topic type. In particular, visual retrieval techniques degrade the performance of topics that are most amenable to textual retrieval techniques. This indicates that systems that can determine the query type may be able to improve performance with that information. However, use of image-based modality classifier can improve the precision of the retrieval, even for tasks that are amenable to textual means.

We obtained moderate results in the ImageCLEFmed automatic annotation task using a neural network approach and primarily low level global features. The best results were obtained by using a feature vector consisting of a 16x16 icon and grey-level co-occurrence features. A multi-layer perceptron architecture was used for the neural network. In the future, we plan to explore using a hierarchical set of classifiers to improve the classification between visually similar classes (for instance, different views of the same anatomical organ). This might also work well with the IRMA classification system.

Acknowledgements

This work was funded in part by NSF Grant ITR-0325160 and NLM Training Grant 1T15 LM009461.

References

- Bishop, C.: *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford (1995)
- Cohen, A., Bhuptiraju, R., Hersh, W.: Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In: *The Thirteenth Text Retrieval Conference: TREC 2004*, Gaithersburg, MD, National Institute of Standards and Technology (2004)
- Cohen, A., Yang, J., Hersh, W.: A comparison of techniques for classification and ad hoc retrieval of biomedical documents. In: *The Fourteenth Text Retrieval Conference - TREC 2005*, Gaithersburg, MD, National Institute for Standards & Technology (2005)
- Crestani, F.: Comparing probabilistic and neural relevance feedback in an interactive information retrieval system. In: *Proceedings of the 1994 IEEE International Conference on Neural Networks*, Orlando, Florida, pp. 3426–3430 (1994)
- Gospodnetic, O., Hatcher, E.: *Lucene in Action*. Manning Publications, Greenwich, CT (2005)
- Han, J., Huang, D.: A novel BP-based image retrieval system. In: *IEEE International Symposium on Circuits and Systems*, Kobe, Japan, 2005, pp. 1557–1560. IEEE, Los Alamitos (2005)

- Haralick, R.: Statistical and structural approaches to texture. *Proceedings of the IEEE* 67, 786–804 (1979)
- Hersh, W., Müller, H., Jensen, J., Yang, J., Gorman, P., Ruch, P.: Advancing biomedical image retrieval: development and analysis of a test collection. *Journal of the American Medical Informatics Association* 13, 488–496 (2006)
- Jensen, J., Hersh, W.: Manual query modification and data fusion for medical image retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, pp. 673–679. Springer, Heidelberg (2006)
- Müller, H., Deselaers, T., Lehmann, T., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: *Evaluation of Multilingual and Multi-modal Information Retrieval - Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain. Springer Lecture Notes in Computer Science, Springer, Heidelberg (in press, 2006)
- Müller, H., Geissbühler, A., Marty, J., Lovis, C., Ruch, P.: The use of MedGIFT and EasyIR for Image CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, pp. 724–732. Springer, Heidelberg (2006)
- Nabney, I.: *Netlab: Algorithms for Pattern Recognition*. Springer, London, England (2004)
- Rahman, M., Desai, B., Bhattacharya, P.: Supervised machine learning based medical image annotation and retrieval in Image CLEFmed 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, pp. 602–701. Springer, Heidelberg (2006)
- Wang, D., Ma, X.: A hybrid image retrieval system with user's relevance feedback using neurocomputing. *Informatica* 29, 271–279 (2005)

MedIC at ImageCLEF 2006: Automatic Image Categorization and Annotation Using Combined Visual Representations

Filip Florea^{1,2}, Alexandrina Rogozan¹, Eugen Barbu¹, Abdelaziz Bensrhair¹,
and Stefan Darmoni^{1,2}

¹ LITIS Laboratory, 1060 Avenue de l'Université
76801 St. Etienne du Rouvray, France

² CISMef Team, Rouen University Hospital & GCSIS
Medical School of Rouen, 1 rue de Germont, 76031 Rouen, France

Abstract. The CISMef group participated at the automatic annotation task of the 2006 ImageCLEF cross-language image retrieval track, employing the MedIC module. The module is designed to automatically extract annotations using image categorization. For the 2006 ImageCLEF annotation experiments we used two sets of visual representations: the first based on the PCA transformation of combined textural and statistic low-level visual features and the second oriented towards compact symbolic image descriptors extracted from the same low-level features. Only the first set of features was present in the actual competition and obtained the fourth rank, with only 1% of error more than the most accurate run of the competition. The comparison with the second representation set was conducted after the official benchmark, on the validation data set, and obtained similar results, but with significantly smaller image signatures.

1 Introduction

When searching for images in non-annotated databases, medical image categorization and annotation can be very useful. In this process, the first step is to define the categories you need to "recognize" and to select a set of image prototypes for each category. Then, each new/unknown image is projected onto one of the categories (using some form of image similarity), and thus annotated with the identifier of the category (or the multilingual terms that can be associated to each category). One of the major disadvantages of this approach is the dependence on manually annotated prototypes for each of the categories, especially when treating domains as vast and rich as medicine. A second weakness is the difficulty to treat images that are not included in the defined categories, an additional category being simply too big to create (basically it should contain everything else). Nevertheless, medical image annotation by categorization proved to be significantly accurate when an image sub-domain is considered, and image categories are well defined.

2 The MedIC Annotation by Categorization Module

The CISMef project¹ (French acronym for Catalog and Index of French-language health resources) [1] is a quality-controlled subject gateway initiated by the Rouen University Hospital. The objective is to describe and index the main French-language health resources (documents on the web) to assist the users (i.e. health professionals, students or general public) in their search for high quality medical information available on the Internet. The CISMef team has developed the MedIC (i.e. Medical Image Categorization) image annotation module, to allow direct access to medical images contained in health-documents. MedIC extracts image annotations using a three-step categorization process:

- (1) the *extraction* of low-level image-feature sets to describe the visual content of medical images
- (2) the *representation* of images by compact visual descriptors.
- (3) the *classification* of the image descriptors according to defined categories.

For the automatic annotation task of the 2006 ImageCLEF cross-language image retrieval track, we used the MedIC module in two configurations. Both are based on the same image features (extracted in step ⟨1⟩) but use different compact representation approaches ⟨2⟩ and classifiers ⟨3⟩. In the following sections we present in detail the methods used at each step.

2.1 Image Resizing, Global and Local Representations

For the 2006 ImageCLEF experiments, all images are resized to a fixed dimension of 256×256 pixels. Even though this solves some of the difficulties raised at the feature extraction step (e.g. simpler generation of texture filters, normalization of the resulted representation size), losing the original image aspect ratio could introduce some structural and textural deformations, which could result in poor performances when dealing, for example, with image fragments. However, from our previous experience and after the examination of the IRMA database we concluded that, most of the times, images of the same category have the same aspect ratio, and thus will finally be deformed in the same way. In applications dealing with image fragments, this resizing step should be avoided.

The image features can be extracted both globally and locally (from image sub-windows). The global features are extracted at the image level and have the advantage of being less sensible to local variations, noise and/or geometrical transformations. However due to the importance of details in medical imaging, local features are of great importance when representing the medical content of images. To capture the spatial disposition of features inside medical images, we choose to extract low-level descriptors from 16 equal non-overlapping image sub-windows (of 64×64 pixels). This way each image is represented by a vector of 16 blocks, and from each block, features are extracted to describe its content.

¹ <http://www.cismef.org>

2.2 Feature Extraction (1)

The X-rays images used for the 2006 ImageCLEF annotation task are by nature acquired in gray-levels, and electronically digitized using 8 bit per pixel ($2^8 = 256$ gray levels). This renders some of the most successfully used (i.e. for image representation) features, like the color, inapplicable in this context. The texture based features combined with statistical gray-level measures proved to be a well suited descriptor for medical images [2].

For describing image *texture* we employed several very different approaches:

- (COOC) From the *co-occurrence matrix* [3] of the four main directions (horizontal, vertical and diagonals), we extract four main measures (after a 64 gray-level quantification): energy, entropy, contrast and homogeneity, producing a **16 feature** vector.
- (FD) Relying on the assumption that textures are fractals for a certain range of magnifications [4], we used a modified *box-counting* texture analysis [5] to compute the *fractal dimension*, obtaining a **single** feature.
- (GB) The belief that simple cells in the human visual cortex can be modeled by Gabor functions [6] lead to the development of texture features extracted from response to *Gabor filters* [7]. Computed at $\lambda = 3$ scales and $\phi = 4$ orientations, we obtain a 12 level filter decomposition from which we extract the mean and standard deviation, resulting a **24 feature** vector.
- (DCT) [8] suggest using a 3×3 *discrete cosine transform* for texture feature extraction. Excluding the low-frequency component of the DCT (poor in textural information), we obtain **8 features**.
- (RL) Galloway has proposed a run-length-based technique, which calculates characteristic textural features from *gray-level run lengths* in different image directions [9]. A total of **14 features** is obtained.
- (Laws) Laws has suggested a set convolution masks for feature extraction [10]. Using the *Laws filter masks for textural energy* **28 features** are resulted.
- (MSAR) [11] proposed the classification of color textures using *Multispectral Simultaneous Autoregressive Model (MSAR)*. The basic idea of a simultaneous autoregressive (SAR) model is to express a gray level of a pixel as a function of the gray levels in its neighborhood. The related model parameters for one image are calculated using a least squares technique and are used as textural features. From this **24 features** are resulting.

In addition we used features derived from *gray-level statistical measures (STAT)*: different estimations of the first order (mean, median and mode), second order (standard deviation and l2norm), third and forth order (skewness and kurtosis) moments, thus obtaining a **7 feature** vector.

In [2], using feature selection algorithms, we show that some of the texture features (i.e. COOC, FD, GB) as well as the statistic features (STAT) are complementary. This was concluded after observing that all ten feature selection methods preserved subsets from each texture feature set. The texture features we added

for the present experiments (i.e. DCT, RL, Laws and MSAR) are behaving similarly, adding a small amount of useful information to the image representation. Therefore we choose to combine all these descriptors to create a richer image representation. However, some of the information is redundant, imposing some form of selection when using combined descriptors.

2.3 Compact Representations ⟨2⟩

As mentioned before, each image is split in 16 equal blocks (sub-images) and for each block a number of $(16+1+24+8+14+28+24+7) = 122$ features are extracted. Considering this approach, concerns related to the size of the learning set are raised. Using a large number of features can lead to poor classifier estimation and an increased computational burden (slow learning and decision time).

To avoid the classification of very rich image representations, we used two methods to transform the image feature vectors into more compact descriptions.

PCA Dimensionality Reduction. In the feature selection comparison conducted in [2], the Principal Component Analysis (PCA) obtained best ratio between dimensionality-reduction and the subsequent classification-accuracy. PCA is a linear transformation that transforms the data into a new coordinate system, the first new coordinate (called the first principal component) containing the projection with the greatest variance of the data (from any projection possible), the second new coordinate containing the second greatest variance and so on. The dimensionality reduction is done retaining those characteristics of the dataset that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Generally, the low-order components often contain the "most important" aspects of the data.

We used this technique for the 2006 ImageCLEF annotation experiments, choosing enough eigenvectors to account for either 95% or 97% of the variance in the original data (on the training set).

Symbolic Description of Images. We propose a second dimensionality reduction approach in the form of a symbolic descriptor. This representation is obtained by assigning a label for every block using a clustering procedure [12].

The first step is to make a crisp partition of the input data set in a number of clusters (400 in the results presented here). The algorithm used is CLARA (Clustering Large Applications) [13]. CLARA finds representative objects, called medoids, in clusters. The second step is to apply a hierarchical ascendant clustering algorithm, AGNES (Agglomerative Nesting) [13], on cluster representants, to add more information to the first results. AGNES use the Single-Link method applied on dissimilarity matrix and merge nodes that have the least dissimilarity in a non-descending fashion.

The hierarchy obtained in the second step is cut, in our application, at four different levels ($C = 100, 200, 300$ and 400 clusters). Every block of the initial

image will be described by one or four labels, leading to either **16** or $4 \times 16 = 64$ characteristics for the initial image. In the recognition stage, each block is labeled using the label of the nearest cluster.

We also employed a second, slightly different, approach for the labeling procedure. In this new setting, each set of blocks sharing the same position in the image decomposition is clustered separately.

2.4 Classification $\langle 3 \rangle$

For the projection of the test instances onto corresponding categories we used two classifiers, one for each type of compact representation. In previous experiments [2] we tested several classifiers (neural networks, decision trees, support vector methods and nearest-neighbor architectures) and noted the best performances obtained by the Support Vector Machine (SVM) classifier and the good performances/classification-time of the k -Nearest Neighbor approach.

Given the SVM's performance when dealing with numerical descriptions, we employed it for classifying the PCA-resulting descriptor. The SVM classification is conducted using different parameters. Using the 1000 images validation dataset, the best parameters are selected: RBF kernel, $\gamma = 10^{-2}$, the cost $C = 10^2$.

For the symbolic descriptors, having to deal with nominal representations, we preferred a k -Nearest Neighbor classifier combined with a VDM (i.e. Value Difference Metric). The VDM metric was introduced by [14] to evaluate the similarity between symbolic (nominal) features.

The k -NN classifier has the advantage to be very fast (e.g. compared to complex architectures, such as SVM), and still very accurate. However, given that in most applications the classification task (especially the learning phase) are expected to be conducted off-line, the classifier's complexity (and therefore the classification speed) is not a major concern.

3 Results

For the Image Annotation Task of the 2006 ImageCLEF our group submitted four runs. The parameters used for each run are presented in Table 1. For all the four runs a combined descriptor **C00C**, **FD**, **GB**, **DCT**, **RL**, **Laws**, **MSAR**, **STAT** is used. This represents $(16+1+24+8+14+28+24+7) = 122$ features for each extracting window (according to section 2.2).

For the **local+global** runs the features are extracted from 1 global window + the 16 local windows, and thus have an initial number of $122 \times 17 = 2074$ features. From these **PCA450** uses a 97% PCA variance to select 450 features, while using 95% produces 335 features. When considering only the 16 local extraction windows, the original number of features becomes $122 \times 16 = 1952$ and the PCA 97% and 95% are producing 333, respectively 150 features.

The **(local+global).PCA335** run obtained the fourth rank, and was situated third in the hierarchy of groups. This run also obtained the third score (the second and third rank having the same error rate: 16.7%). The best score was

Table 1. Submitted run details (PCA)

Run label	(a)					(b)		
	Parameters					error	error	rank
	local	global	initial	PCA	final	error	error	
			no.feet.	var	no.feet.	valid.	test	
local_PCA333	✓	×	1952	97%	333	12.6%	17.2%	5
local_PCA150	✓	×	1952	95%	150	15.9%	20.2%	10
(local+global)_PCA450	✓	✓	2074	97%	450	12.3%	17.9%	7
(local+global)_PCA335	✓	✓	2074	95%	335	12.7%	17.2%	4

obtained by the RWTHi6 group with an error rate of 16.2%. That represents an improvement of 1% compared to our best score (local+global)_PCA335, meaning that, compared to our 828 correctly annotated test images, 10 more images (not necessarily from those we missed) are correctly annotated.

However, it is interesting to note that the validation-error is better with (in average) 4.75% than the test-error, reaching up to 87.7% correctly annotated validation images (12.3% error rate) for local+global_PCA450. This indicates different difficulties for the validation and test sets, and it will be interesting to compare if different systems reacted in the same way.

Also, we note that equal test error rates are obtained for two runs: local_PCA333 and (local+global)_PCA335, with reduced representations of comparable sizes. This indicates that the information added by the global representation is redundant. Using the same parameters, we obtain similar performances on the validation dataset.

The second dimensionality reduction approach was evaluated on the validation dataset, and compared with the result obtained on the same dataset by the PCA-based approach. The results, in terms of classification error are presented in the Table 2.

We tested this approach using the 1NN classifier and the position-dependent block clustering, as we showed that they perform better [12]. The Table 2 shows the performances obtained using symbolic representations at four levels of complexity ($C = 100, 200, 300$ and 400 clusters) as well as a representation using all four levels combined (4-level). Using the symbolic representation we obtain, four representations of 1×16 features and one of $4 \times 16 = 64$ features, when using only the local description (16 blocks), and four representations of 1×17 features and one of $4 \times 17 = 68$ features when the global description is added.

The two symbolic approaches are producing similar results, when evaluating the accuracy/no. of features ratio (accuracy=100%-error). However, similar to the previous method, if the processing time is not an important issue, using both local and global descriptors is preferable, as this obtains slightly better results.

The differences between representations using 100 clusters and 400 clusters are never bigger than 1%. This indicates that the proposed symbolic representation captures similar image information at different levels. This assumption is verified by the small improvement obtained when joining the representation vectors (the

Table 2. Symbolic run

Run label	(a)				(b)		
	Parameters				error valid.	error test	rank
	local	global	initial no.feats.	final no.feats.			
local_symbol100	✓	×	1952	16	15.1%	-	-
local_symbol1200	✓	×	1952	16	14.9%	-	-
local_symbol1300	✓	×	1952	16	14.5%	-	-
local_symbol1400	✓	×	1952	16	14.4%	-	-
local_symbol14-level	✓	×	1952	(4*16)=64	13.9%	-	-
(local+global)_symbol100	✓	✓	2074	17	14.8%	-	-
(local+global)_symbol1200	✓	✓	2074	17	14.8%	-	-
(local+global)_symbol1300	✓	✓	2074	17	14.5%	-	-
(local+global)_symbol1400	✓	✓	2074	17	14.1%	-	-
(local+global)_symbol14-level	✓	✓	2074	(4*17)=68	13.7%	-	-

4-level set). Therefore, using a representation using all four symbolic resolutions increases significantly (four times) the final feature-space dimensionality and only slightly enhances the capacity of capturing the image content.

Compared to the PCA-based approach, the symbolic representation of the image content obtains inferior results, but using a significantly smaller representation space. As a comparison, the `local.PCA150` set obtains weaker results (15.9% error of validation set) than any of the symbolic sets, using a representation space significantly bigger.

In a number of applications it would be useful to conserve not only the category and the annotation extracted by an architecture like the one we presented, but also the actual image (numeric or symbolic) description (signature), to be used later for another operation (such as a visual similarity query). A symbolic approach can be better suited for such a context due to its compactness and the facility of evaluating nominal similarity with VDM.

In a previous experiment [12], we compared the two dimensionality reduction methods (PCA vs. symbolic) on another database of comparable size. We use an image database consisting of 10322 images divided into 33 classes. The images are representing six medical-imaging modalities, each containing a hierarchy of anatomical (e.g. head, thorax, lower-leg), sub-anatomical regions (e.g. knee, tibia, ankle) and acquisition-views (coronal, axial, sagittal). When limited at the same dimensionality (**16** and $4*16=64$), the symbolic descriptor performed better than the PCA-based method.

4 Conclusion

In this paper we present the methods we used for the ImageCLEF 2006 automatic annotation evaluation. From the four runs we submitted, the highest accuracy was achieved by the `(local+global).PCA335` run, thus obtaining the fourth rank

(also the third score and the third placed group). A second symbolic approach was tested outside the official benchmark, and showed comparable (yet slightly inferior) results, but using significantly smaller image representations. A richer dictionary of symbols (more clustering classes) describing smaller image blocks is believed to improve the performances.

The results obtained in the annotation task show that the approach we propose is capable of obtaining very competitive results. A comparison between the correctly annotated images of different systems could be interesting, and could indicate how to combine different architectures to further improve the categorization/annotation results.

References

1. Darmoni, S., Leroy, J., Thirion, B., Baudic, F., Douy  re, M., Piot, J.: Cismef: a structured health resource guide. *Methods of Information in Medicine* 39(1), 30–35 (2000)
2. Florea, F., Rogozan, A., Bensrhair, A., Darmoni, S.: Comparison of feature-selection and classification techniques for medical image modality categorization. In: 10th IEEE Conference OPTIM, vol. 4, pp. 161–168 (2006)
3. Haralick, R.M., Shanmugam, K., Dinstein, I.: Texture features for image classification. *IEEE Trans. Systems, Mans and Cybernetics SMC-3*, 610–621 (1973)
4. Pentland, A.: Fractal-based descriptors of natural scenes. *IEEE Trans. on PAMI* 6(6), 661–674 (1984)
5. Keller, J., Chen, S., Crownover, R.: Texture description and segmentation through fractal geometry. *CVGIP* 45, 150–166 (1989)
6. Marcelja, S.: Mathematical description of the response of simple cortical cells. *Journal of Optical Society of America* 70, 1297–1300 (1980)
7. Lee, T.: Image representation using 2d gabor wavelets. *IEEE Trans. on PAMI* 18(10), 1–13 (1996)
8. Ngo, C.W., Pong, T.C., Chin, R.T.: Exploiting image indexing techniques in DCT domain. *Pattern Recognition* 34(9), 1841–1851 (2001)
9. Galloway, M.M.: Texture analysis using graylevel runlengths. *Computer, Graphics and Image Processing* 4, 172–179 (1975)
10. Laws, K.: Textured Image Segmentation. PhD thesis, University of Southern California (1980)
11. Mao, J., Jain, A.: Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition* 5(2), 173–188 (1992)
12. Florea, F., Barbu, E., Rogozan, A., Bensrhair, A., Buzuloiu, V.: Medical image categorization using a texture based symbolic description. In: 13th International Conference on Image Processing ICIP, Atlanta, USA, pp. 1489–1492 (2006)
13. Kaufman, L.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
14. Stanfill, C., Waltz, D.: Toward memory based reasoning. *Communications of the ACM* 29(12), 1213–1228 (1986)

Medical Image Annotation and Retrieval Using Visual Features

Jing Liu^{1,*}, Yang Hu^{2,*}, Mingjing Li³, Songde Ma¹, and Wei-ying Ma³

¹ Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
jliu@nlpr.ia.ac.cn, mostma@gmail.com

² University of Science and Technology of China, Hefei 230027, China
yanghu@ustc.edu

³ Microsoft Research Asia, No 49, Zhichun Road, Beijing 100080, China
{mjli, wyma}@microsoft.com

Abstract. In this paper, we present the algorithms and results of our participation in the medical image annotation and retrieval tasks of ImageCLEFmed 2006. We exploit using global features and local features to describe medical images in the annotation task. Different kinds of global features are examined and the most descriptive ones are extracted to represent the radiographs, which effectively capture the intensity, texture and shape characters of the image content. We also evaluate the descriptive power of local features, i.e. local image patches, for medical images. A newly developed spatial pyramid matching algorithm is applied to measure the similarity between images represented by sets of local features. Both descriptors use multi-class SVM to classify the images. We achieve an error rate of 17.6% for global descriptor and 18.2% for the local one, which rank sixth and ninth respectively among all the submissions. For the medical image retrieval task, we only use visual features to describe the images. No textual information is considered. Different features are used to describe gray images and color images. Our submission achieves a mean average precision (MAP) of 0.0681, which ranks second in the 11 runs that also only use visual features.

1 Introduction

Due to the rapid development of biomedical informatics, medical images have become an indispensable investigation tool for medical diagnosis and therapy. The ever-increasing amount of digitally produced images require efficient methods to archive and access this data. Therefore, the application of general image classification and retrieval techniques in this specialized domain has obtained increasing research interest recently.

In this paper, we describe our participation in the automatic medical image annotation and medical image retrieval tasks of ImageCLEF 2006. We submitted two runs for the annotation task, which exploited, respectively, the effectiveness

* This work was performed when the first and the second authors were visiting students at Microsoft Research Asia.

of global features and local features for medical image classification. We also investigated using visual features only for medical image retrieval. A detailed description of the tasks and the databases, as well as an overall analysis of the submissions and results from all participants can be found in [7].

2 Automatic Medical Image Annotation

The automatic image annotation task is to classify images into a set of predefined categories. It provides a dataset of 10,000 images, which are classified into 116 categories, for participants to train a classification system. 1000 additional radiographs are used to evaluate the performance of various algorithms.

We developed two different schemes for this task. In the first algorithm, traditional global features, such as intensity, texture and shape descriptors were used to describe medical images. In the second one, we exploited using local features to represent the images. And a spatial pyramid matching scheme was then applied to measure the similarity between two images. Both methods used SVM to classify the images into different categories.

2.1 Global Features for Medical Image Classification

When designing image features, we should consider two issues. First, the features should be representative for the images. Second, the complexity of calculating the features should be relatively low. Medical images have their particular characteristics in appearance. For example, radiographs are usually grayscale images and the spatial layouts of the anatomical structures in the radiographs of the same category are quite similar. Therefore, the texture, shape and local features should be valuable and discriminative for describing radiographs.

According to these observations, we select several different kinds of visual features to represent the radiographs. We extract gray-block feature and block wavelet feature from the original images. Shape-related features are extracted from the corresponding binary images. Then, they are combined into a 382-dimensional feature vector. The detail descriptions of the features are as follows:

Gray-Block Feature. The original images are uniformly divided into $8 \times 8 = 64$ blocks. The average gray value in each block is calculated and a 64-dimensional gray-block feature is obtained. The ℓ_2 -norm of the feature vector is set to 1. The normalization could reduce the influence of illumination variance across different images to some extent. According to the experiments, this is the most effective feature although it is straight forward and very simple.

Block-Wavelet Feature. The wavelet coefficients could characterize the texture of the images at different scales. We divide the images into $4 \times 4 = 16$ blocks and extract multi-scale wavelet features in each block. We implement 3-level wavelet transforms on the image blocks using Daubechies filter (db8). Then, the mean and the variance of the wavelet coefficients in the HL, LH

and HH sub-bands are computed. Therefore, we get a $288(6 \times 3 \times 4 \times 4)$ -dimensional feature vector.

Features for the Binary Image. We first convert the images into binary images through thresholding. Otsu's method is used here to calculate the threshold. The area and the center point of the object region in the binary image are calculated. Moreover, we apply morphological operations on the binary image and extract the contour and the edges of the object. The length of the contour and the ratio of the total length of the edges and that of the contour are calculated and are taken as the shape feature. Then we get a 5-dimensional feature for the binary image. Although the dimension of this feature is small, it is highly discriminative among different categories. In order to increase the effect of this feature, we duplicate it 6 times and convert it into a 30-dimensional feature vector.

Choosing suitable parameters for above features is very difficult in theory. Therefore, we tune the parameters through experiments. The parameters, such as the size of the image block and the dimension of the features for the binary image, are determined through cross-validation on the training set. The same parameter settings are used in both of the annotation task and the retrieval task.

The classifier is trained using SVM, which is a classic machine learning technique that has strong theoretical foundation and excellent empirical successes. The basic idea of SVM is to map the data into a high dimensional space and then find a separating hyperplane with the maximal margin. In the experiment, we use the multi-class SVM implemented by the LIBSVM tool[1]. The radial basis function (RBF) is chosen as the kernel function and the optimal parameters are determined through 5-fold cross-validation.

2.2 Local Features and Spatial Pyramid Matching

Recently, a class of local descriptor based methods, which represent an image with an collection of local photometric descriptors, have demonstrated impressive level of performance for object recognition and classification. And this kind of algorithms have also been explored for medical image classification, considering that most information in medical images is local. Many recent works have devoted to leverage the power of both local descriptor and discriminative learning. Lazebnik et al.[5] presented a spatial pyramid matching method for recognizing natural scene categories. It constructed a pyramid in image space by partitioning the image into increasingly fine sub-regions. The histograms of local features were computed on each sub-region and a weighted histogram intersection was applied to measure the similarity between feature sets. Because the geometric information of local features is extremely valuable for medical images, in which the objects are always centered in the images and the spatial layouts of the anatomical structures in the radiographs belonging to the same category are quite similar, we can expect promising results using this spatial matching scheme. We apply spatial pyramid matching for medical image classification and examine its performance on this new task.

Although SIFT descriptor [6] has been proven to work well for common object and nature scene recognition [5], its power to describe radiographs is somewhat limited. Since the scale and rotation variations in radiographs of the same category are small, the SIFT descriptor can not show its advantage of being scale and rotation invariant for describing radiographs. In previous works, local image patches have shown pleasant performance for medical image retrieval and classification [4][2]. Therefore, we utilize local image patches as the local features in our experiments. Before feature extraction, we resize the images so that the long sides are 200 pixels and their aspect ratios are maintained. The positions of the local patches are determined in two ways. Local patches are first extracted from interest points detected by DoG region detector [6], which are located at local scale-space maxima of the Difference-of-Gaussian. We also extract local patches from an uniform grid spacing at 10×10 pixels. This dense regular description is necessary to capture uniform regions that are prevalent in radiographs. We use 11×11 pixel patches in our experiments. And about 400 patches are extracted from each image. After feature extraction, we applied a high speed clustering algorithm Growing Cell Structures (GCS) neural network [3], which is able to detect high dimensional patterns with any probability distribution, to quantize all feature vectors into M discrete types ($M = 600$ in the experiment). Then each feature vector is represented by the ID of the cluster it belongs to and its spatial coordinate.

In order to measure the similarity between two images represented by orderless collections of local patches, we first partition the scaled images into increasingly fine sub-regions. Then we compute the histograms of cluster frequencies inside each sub-region by counting the number of patches that belong to each cluster (Fig. 1). The histograms from two images are compared using a weighted histogram intersection measure. Let X and Y be two sets of feature vectors representing two images. Their histograms in the i th sub-region at level l are denoted by H_X^{li} and H_Y^{li} . $H_X^{li}(j)$ and $H_Y^{li}(j)$ indicate the number of feature vectors that fall into the j th bin of the histograms. The similarity between X and Y is defined as follows:

$$\mathcal{K}(X, Y) = \sum_{l=1}^L w_l \sum_{i=1}^{4^{(l-1)}} \sum_{j=1}^M \min(H_X^{li}(j), H_Y^{li}(j)) , \quad (1)$$

where L refers to the max level. As shown in Fig. 1, the weight w_l is inversely proportional to region size: the smaller the region the larger the weight, i.e. matches made within smaller regions are weighted more than those made in larger regions. The histogram intersection function $\min(H_X^{li}(j), H_Y^{li}(j))$ measures the “overlap” between two histograms’ bins, which implicitly finds the correspondences between feature vectors falling into that sub-region.

$\mathcal{K}$ has been proved to satisfy the Mercer’s condition, i.e. it is positive semi-definite [5]. Therefore, kernel-based discriminative methods can be applied. In the experiment, multi-class classification is done with a “one-against-one” SVM classifier [1] using the spatial pyramid matching kernel.

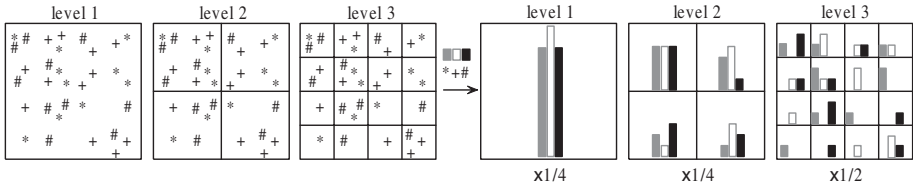


Fig. 1. Toy example of constructing a three-level spatial pyramid. The image has three types of features, indicated by asterisks, crosses and pounds. At the left side, the image is subdivided at three different levels of resolution. At the right, the number of features that fall in each sub-region is counted. The spatial histograms are weighted during matching [5].

3 Medical Image Retrieval

The dataset for the medical image retrieval task consists of images from the Casimage, MIR, PEIR and PathoPIC datasets. There are totally 50,026 images with different modalities, such as photographs, radiographs, ultrasonic images, and scans of illustrations used for teaching etc. Query topics are formulated with example images and a short textual description, which denotes the exact information need such as the illness, the body region or the modalities shown in the images. Therefore, this task is much more challenging than the annotation task. We only exploit the effectiveness of visual features for this task. No textual information is utilized in our experiment.

As general image retrieval systems, the whole retrieval procedure contains three steps: image preprocessing, feature extraction and relevance ranking based on similarity measure. For image preprocessing, we first resize the images so that the long sides are 512 pixels and their aspect ratios are maintained. As the characters of gray images and color images are quite different, we examine whether an image is gray or color before extracting features from it. Feature extraction is carried out according to the type of the image, i.e. the features for gray image and color image are different:

Features for Gray Images. The global features used to describe radiographs in the annotation task are used here to describe the gray images (see Section 2.1).

Features for Color Images. We use band-correlogram, color histogram and block-wavelet features to describe the color images:

- *Band-Correlogram.* We first quantize the RGB values into 64 bins. Then the general auto-correlogram features are extracted within four square neighborhoods, whose radius are 1,3,5,7 pixels respectively. The final features used are the average of the corresponding elements in the four square neighborhoods. It is a 64-dimensional feature vector.
- *Color Histogram.* We quantize the RGB values into 36 bins, and calculate the 36-dimensional color histogram.

- *Block-Wavelet*. We convert the color images into gray images and calculate the block-wavelet feature as introduced in Section 2.1.

The last step is ranking the images in the dataset according to their relevance to the query images. As each topic contains multiple query images, the distance between a dataset image S and a set of query images belonging to the same topic is defined as the minimum distance between S and each query image:

$$d(S, Q) = \min_i d(S, Q_i) . \quad (2)$$

The top 1000 images are returned for evaluation.

4 Experimental Results

4.1 Results of Automatic Medical Image Annotation

For the annotation task, we submitted two runs named “msra_wsm_gray” and “msra_wsm_patch”, which correspond to global feature and local feature methods respectively. The submission using global features achieved an error rate of 17.6%, which ranked sixth among all the submissions. And the error rate of the run using local features is 18.2%, which ranked ninth.

In order to examine whether these two methods are complementary to each other for describing radiographs, we combine them and test the new classifier on the test dataset. We first calculate the probabilistic outputs of the two SVMs. Let $P_G(c|x)$ and $P_L(c|x)$ correspond to the outputs of SVMs using global feature and local feature respectively, where c represents the class and x is the test image. The combined posteriori probability is calculated as follows:

$$P(c|x) = \frac{1}{Z} [w_G^c P_G(c|x) + w_L^c P_L(c|x)] , \quad (3)$$

where $0 \leq w_G^c, w_L^c \leq 1$, $w_G^c + w_L^c = 1$ and Z is a normalization constant. w_G^c and w_L^c are determined according to the relative performance of the two classifiers for class c on the training set. The classifier with better performance on class c is weighted heavier. We name the combined classifier as CombiSVM. It achieves an error rate of 16.2% on the test set, which equals the best result of all the submissions this year.

Fig. 2 illustrates the performance of CombiSVM for each category on the test dataset. Through analyzing these results, we can see that CombiSVM achieves satisfactory precision on most categories. Specifically, 40 categories achieve 100% precision. Bad classification performance generally happens on categories with relatively small number of training images. The categories with zero precision always have training images less than 20.

4.2 Results of Medical Image Retrieval

In the medical image retrieval task, the parameters for gray images are the same with the annotation task. The parameters for color images are determined

Table 1. Error rates for the medical automatic annotation task

<i>Rank</i>	<i>Method</i>	<i>Error rate</i> [%]
-	CombiSVM	16.2
1	SHME	16.2
6	msra_wsm_gray	17.6
9	msra_wsm_patch	18.2
-	baseline (medGIFT)	32.0

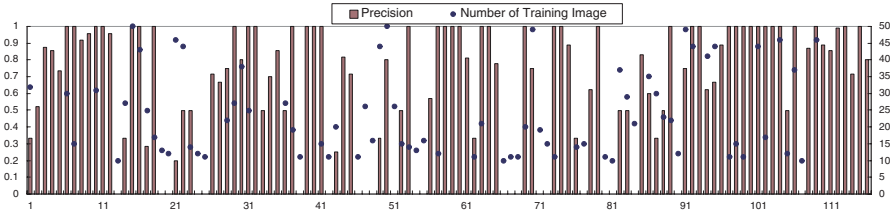


Fig. 2. Classification precision of CombiSVM for each category on the test set: the precisions are represented as bars according to the left axis. The numbers of training images for categories, which have less than 50 training images, are shown as points according to the right axis.

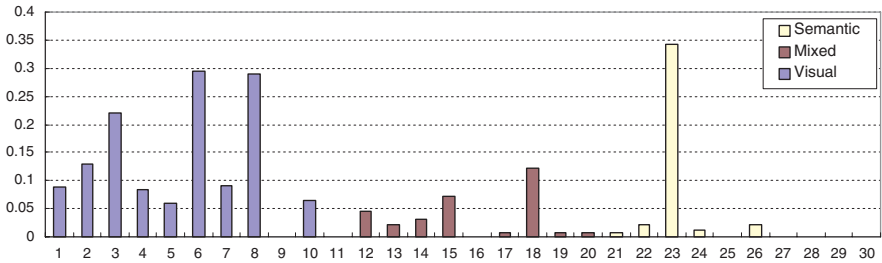


Fig. 3. Mean average precision for each query topic

empirically. The details have been discussed in Section 3. We employ these features in our automatic “visual retrieval” system and submit only one run named “msra_wsm”. We achieved a MAP of 0.0681, which ranks second among the 11 runs that also only use visual features. The MAP of the best run is 0.0753.

The MAP values for each query are shown in Fig. 3. We use different color bars to indicate the different performances on visual, mixed and semantic topics. The average MAP on these three kinds of topics are 0.1324, 0.0313 and 0.0406 respectively. It is obvious that the performance on visual topics is best. The performance is relatively poor on other topics with more semantic considerations. The differences between the performances on different kinds of topics are reasonable considering the design of the topics. The MAP for the 23rd topic which is a

semantic topic is strangely high. It is because there are large number of images that are extremely similar with the query images of this topic.

5 Conclusion

In this paper, we present our work on the medical image annotation and retrieval tasks of ImageCLEFmed 2006. Due to the special characteristics of radiographs, we explored using global and local features respectively to describe them in the annotation task. Then we use the multi-class SVM to classify the images. We achieved an error rate of 17.6% for the global feature based method and 18.2% for the local feature method. For the medical image retrieval task, we distinguished gray images from color images and used different kinds of visual features to describe them. Our submission ranked second among the 11 runs which also only used visual features.

This is our first participation in the tasks concerning medical images. We find this task quite interesting and very challenging. In our future work, we will investigate some more descriptive features and more suitable similarity measure for comparing images. We didn't utilize textual information in our experiment. We will incorporate it into the retrieval framework in the future.

References

1. Chang, C.-C., Lin, C.-J.: LIBSVM : A Library for Support Vector Machines Software (2001), available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
2. Deselaers, T., Keysers, D., Ney, H.: Discriminative Training for Object Recognition Using Image Patches. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA (June 2005)
3. Fritzke, B.: Growing Cell Structures - A Self-Organizing Network in k Dimensions. *Artificial Neural Networks II*, 1051–1056 (1992)
4. Keysers, D., Gollan, C., Ney, H.: Classification of Medical Images using Non-linear Distortion Models. *Bildverarbeitung für die Medizin 2004 (BVM 2004)*, Berlin, Germany, pp. 366–370 (March 2004)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York (June 2006)
6. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
7. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain (in press, 2006)

Baseline Results for the ImageCLEF 2006 Medical Automatic Annotation Task

Mark O Güld, Christian Thies, Benedikt Fischer, and Thomas M Deserno

Department of Medical Informatics, RWTH Aachen, Aachen, Germany
{mguel,d,cthies,bfischer}@mi.rwth-aachen.de, lehmann@computer.org

Abstract. The ImageCLEF 2006 medical automatic annotation task encompasses 11,000 images from 116 categories, compared to 57 categories for 10,000 images of the similar task in 2005. As a baseline for comparison, a run using the same classifiers with the identical parameterization as in 2005 is submitted. In addition, the parameterization of the classifier was optimized according to the 9,000/1,000 split of the 2006 training data. In particular, texture-based classifiers are combined in parallel with classifiers, which use spatial intensity information to model common variabilities among medical images. However, all individual classifiers are based on global features, i.e. one feature vector describes the entire image. The parameterization from 2005 yields an error rate of 21.7%, which ranks 13th among the 28 submissions. The optimized classifier yields 21.4% error rate (rank 12), which is insignificantly better.

1 Introduction

The ImageCLEF medical automatic annotation task was established in 2005 [1], demanding the classification of 1,000 radiographs into 57 categories based on 9,000 categorized reference images. The ImageCLEF 2006 annotation task [2] consists of 10,000 reference images grouped into 116 categories and 1,000 images to be automatically categorized. This paper aims at providing a baseline for comparison of the experiments in 2005 and 2006 rather than presenting an optimal classifier.

2 Methods

According to our 2005 submission, global features are used to represent the texture of medical images, i.e., one feature vector of about 1,000 bins represents the entire image [3].

2.1 Image Features

The content of a medical image is represented by texture features proposed by TAMURA ET AL. [4] and CASTELLI ET AL. [5], denoted as TTM and CTM, respectively. Down-scaled representations of the original images were computed to

32×32 and $X \times 32$ pixels disregarding and according to the original aspect ratio, respectively. Since these image icons maintain the spatial intensity information, variabilities which are commonly found in a medical imagery are modelled by the distance measure. These include radiation dose, global translation, and local deformation. In particular, the cross-correlation function (CCF) which is based on SHANNON and the image distortion model (IDM) suggested by KEYSERS ET AL. [6] is used. In particular, the following parameters were set:

- **TTM**: texture histograms, 384 bins, Jensen-Shannon divergence
- **CTM**: texture features, 43 bins, Mahalanobis distance with diagonal covariance matrix Σ .
- **CCF**: 32×32 icon, 9×9 translation window
- **IDM**: $X \times 32$ icon, gradients, 5×5 window, 3×3 context

2.2 Classifiers

The single classifiers are combined within a parallel scheme, which performs a weighting of the normalized distances obtained from the single classifiers C_i , and applies the nearest-neighbor-decision function C to the resulting distances:

$$d_c(q, r) = \sum_i \lambda_i \cdot d_i(q, r) \quad (1)$$

where λ_i , $0 \leq \lambda_i \leq 1$, $\sum_i \lambda_i = 1$ denotes the weight for the normalized distance $d_i(q, r)$ obtained from classifier C_i for a sample q and a reference r .

2.3 Submissions

Based on a combination of classifiers used for the annotation task in 2005 [7], a run using the exact same parameterization is submitted. Additionally, a second run is submitted which uses a parameterization obtained from an exhaustive search (using a step size of 0.05 for λ_i) for the best combination of the single classifiers. For this purpose, the development set of 1,000 images is used and the system was trained on the remaining 9,000 images.

3 Results

All results are obtained non-interactively, i.e. without relevance feedback by a human user. Table 1 shows the error rates in percent obtained for the 1,000 unknown images using single k -nearest neighbor classifiers and their combination for both $k = 1$ and $k = 5$. The error rate of 21.4% ranks 12th among 28 submitted runs for this task. The weights that were optimized on the ImageCLEF 2005 medical automatic annotation task (10,000 images from 57 categories) yield an error rate of 21.7% and rank 13th.

Table 1. Error rates for the medical automatic annotation task

Classifier	λ_{TMM}	λ_{CTM}	λ_{CCF}	λ_{IDM}	$k = 1$	$k = 5$
TMM	1.00	0	0	0	44.4%	44.9%
CTM	0	1.00	0	0	27.9%	25.7%
CCF	0	0	1.00	0	23.0%	23.4%
IDM	0	0	0	1.00	57.2%	54.9%
ImageCLEF 2005	0.40	0	0.18	0.42	21.7%	22.0%
Exhaustive search	0.25	0.05	0.25	0.45	21.5%	21.4%

4 Discussion

The comparability of similar experiments such as ImageCLEF 2005 and ImageCLEF 2006 is difficult since several parameters such as the images and the class definitions were changed. In general, the difficulty of a classification problem is proportional to the error rate, increases with the number of classes, and decreases with the number of references per class. Since the number of references varies for the classes, this relation is rather complex than linear. However, linearity can be used as a rough but first approximation.

Since the error rates are not substantially improved by a training on the 2006 data, the classifier that has been optimized for 2005 is also suitable for the ImageCLEF 2006 medical automatic annotation task. The parameterization optimized for ImageCLEF 2005 and 2006 yields $E^{2005} = 13.3\%$ and $E^{2006} = 21.7\%$, respectively, and the quotient E^{2005}/E^{2006} can be used to relate the results from the 2005 and 2006 campaigns. In other words, the best $E = 16.2\%$ from 2006 would approximately have resulted in $E = 10.0\%$, which is better than the best rate obtained in 2005. This is in accordance with the rank of submission, that drops from the 2nd to the 13th place. In conclusion, general improvements for automatic image annotation have been made during the last year.

Acknowledgment

This work is part of the IRMA project, which is funded by the German Research Foundation, grant Le 1108/4.

References

1. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
2. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: CLEF Working Notes, Alicante, Spain (inpress, 2006)

3. Lehmann, T.M., Güld, M.O., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.B.: Content-based image retrieval in medical applications. *Methods of Information in Medicine* 43(4), 354–361 (2004)
4. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics* 8(6), 460–473 (1978)
5. Castelli, V., Bergman, L.D., Kontoyiannis, I., Li, C.S., Robinson, J.T., Turek, J.J.: Progressive search and retrieval in large image archives. *IBM Journal of Research and Development* 42(2), 253–268 (1998)
6. Keysers, D., Dahmen, J., Ney, H., Wein, B.B., Lehmann, T.M.A.: statistical framework for model-based image retrieval in medical applications. *Journal of Electronic Imaging* 12(1), 59–68 (2003)
7. Güld, M.O., Thies, C., Fischer, B., Lehmann, T.M.: Content-based retrieval of medical images by combining global features. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, pp. 702–711. Springer, Heidelberg (2006)

A Refined SVM Applied in Medical Image Annotation

Bo Qiu

Institute for Infocomm Research (I²R)
21 Heng Mui Keng Terrace,
119613, Singapore
qiubo@i2r.a-star.edu.sg

Abstract. For the medical image annotation task of ImageCLEF2006, we developed a refined SVM method. This method includes two stages, in which accordingly the coarse and fine classifications are performed. At the coarse stage, the common Support Vector Machines (SVM) method is used with the corresponding image feature, down-scaled low resolution pixel map (32x32). At the refined stage, the results from the first stage are refined following the steps: 1) select those images near to the borders of SVM classifiers, which are to be re-classified and selected by a predefined threshold of SVM similarity value; 2) form a new training dataset, to eliminate the influence of the great volume unbalance among classes; 3) use three classification methods and image features: 20x50 low resolution pixel maps feed into SVMs; SIFT features feed into Euclidean distance classifiers; 16x16 low resolution pixel maps feed into PCA classifiers. 4) combine the results from 3). At last the results from the two stages are combined to form the final classification result. Our experimental results showed that with the two-stage method a significant improvement of the classification rate has been achieved.

Keywords: Medical image annotation, refined SVM.

1 Introduction

In the task the annotation task is simplified into a classification problem: purely using visual features to classify some unlabeled images into some defined classes which are given by doctors. The total process is demanded to be automatic.

In ImageCLEF2006 there are 116 classes to be classified, which is totally different from 57 classes in ImageCLEF2005. Another difference is that there is an extra DEV set including 1000 labeled images. The common parts of 2005 and 2006 are: 9000 labeled training images, 1000 testing images. And the evaluation method is simple: to minimize the number of misclassified images in the 1000 testing data.

In general each solution for the classification problem will include how to choose image features and how to construct the corresponding classifiers. By now, many features and classifiers have been explored. However, finding the most efficient combination of the features and the classifier is the key factor to the solution.

2 Image Features

According to [1], the most frequently used image features are categorized into color, texture, and shape. In our past work [2], we tested Low Resolution Pixel-Map (LRPM) which can be regarded as a kind of color feature; beside of it, we also explored contrast, anisotropy, polarity, which are texture features, and Blob, a mid-level feature compared with the common used low-level features.

In this year, we tested some new features (as shown in Fig.1) like salient map [3], salient point [4], SIFT [5], and stripe [6].

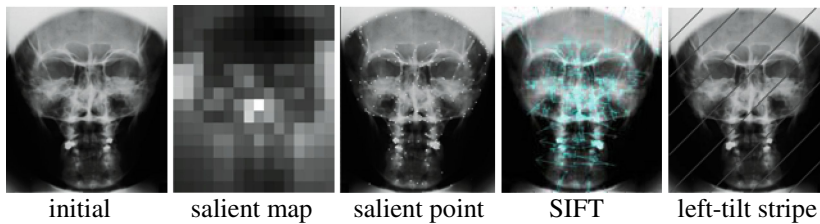


Fig. 1. Features tested in this year

Facing so many features, we can not judge which one or which combination is the best for the solution of the task. The only way is to choose them by experiments. Besides of the features, another important factor which greatly influences the classification results is the choice/design of classifiers.

3 Classification Method: Coarse-to-Fine SVM

According to [7], there have been many classifiers developed, like Boosting, Decision trees, Neural Networks, Bayesian Networks, SVM, Hidden Markov Field, etc. In our system, SVM is chosen as the classifier due to its efficient ability to handle a multi-class classification problem, which has been proven by our former work and widely used in ImageCLEF work groups [8].

3.1 A Basic Classification System Using SVM (Coarse)

As shown in Fig.2, SVM is a supervised method. So the learning process is needed, in which the main target is to tune the parameters. After the parameters are fixed by the cross validation part, the unlabeled images can be processed by the classification system.

For each input image, there will be a distance vector whose elements represent the distances from the image to the boundaries of each class. This distance vector is generated from the SVM classifier, and is called similarity vector. Obviously, the smaller the similarity value is, the lower the possibility is that the image is belonging to the class. The largest value ‘absorbs’ the image into the corresponding class.

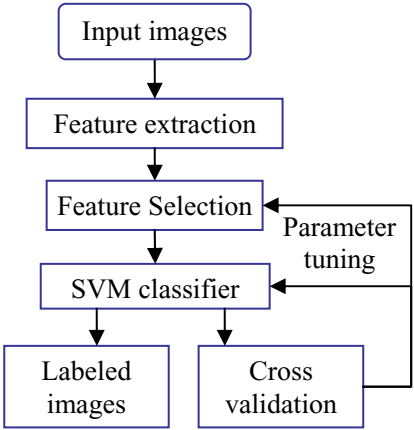


Fig. 2. A basic classification system

3.2 Refined SVM (Fine)

The basic SVM as shown in section 3.1 can be regarded as the coarse part of our two-stage SVM. The second part will be how to find the badly classified classes from the first stage and refine the results. This is based on the cross validation part. For one image in those badly-classified classes, note that its corresponding distance vector is $D = (d_1, d_2, d_3, \dots, d_n)$. Inside D 's n elements, the maximal 3 values are d_{k1}, d_{k2}, d_{k3} . Then if

$$(d_{k1} - d_{k2}) \times 10000 \leq th_1, \quad (1)$$

which hints the image is probably belonging to class $k2$ instead of $k1$, it should be reclassified. Threshold th_1 is decided by cross validation.

However, for all the images chosen by this way, there are around 1/3 of them are correctly classified, which has been proved by cross validation. This means the 1/3 needn't to be reclassified. How to identify those images? We find that if the image distance vector satisfies the following condition

$$(d_{k2} - d_{k3}) \times 10000 > th_2, \quad (2)$$

the image will be regarded as a 'good' image and do not need to be reclassified. th_2 is also a threshold decided by cross validation.

While keeping the classified results for all the 'good' images, we start to reclassify the 'bad' images. The process includes the following steps:

- 1) Generating a new training dataset out of the old one. This is to eliminate the influence of the great volume unbalance among classes.
- 2) Feeding 20x50 LRPM into SVM classifiers; Feeding SIFT features into Euclidean distance classifiers; Feeding 16x16 LRPM into PCA classifiers;
- 3) Choosing the best classifier and features for each class using cross validation; then classifying the 'bad' images and combine the results from the 'good' images.

4 Application: ImageCLEF'06 Medical Annotation

In the case of fulfilling the task of ImageCLEF2006, we use the DEV set to make the cross validation to train the classifiers and choose image features. Then the 1000 true images are tested based on the trained classifiers. When using the coarse-to-fine SVM, our system reached an average accuracy of 78.2% for the DEV set, compared to the 70.1% purely derived from the basic SVM system. In the whole process of cross validation, the SVM parameters are set and the two thresholds in formula (1) and (2) are chosen: $th_1 = 1.37$, $th_2 = 3$.

Using the same SVM parameters and thresholds, when applying our method into the true test dataset, the final classification accuracy is 72%.

5 Conclusion and Future Work

Though we got a comparable result for the DEV set, the true result with the 72% accuracy is a bit far away from expected. The main reason for this may be due to the choice of features. Global features are not good enough to be applied to solve the complex classification problem of medical images. In ImageCLEF's working notes we can see that more precise local features have to be considered. For future work, we will focus on more efficient features and relevant classifiers. And more work on analyzing the 116 classes should be done because lots of classes cannot be identified even with our own eyes. Where are the key parts to distinguish them? It may be a long-term question.

References

- [1] Lehmann, T., et al.: Automatic categorization of medical images for content-based retrieval and data mining. *Computerized Medical Imaging and Graphics* 29, 143–155 (2005)
- [2] Qiu, B., Xu, C.S., Tian, Q.: An automatic classification system applied in medical images. In: *IEEE Intl. Conf. Multimedia & Expo (ICME) 2006*, Toronto, Canada, July 9-12, 2006 (2006)
- [3] Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research* 40(10-12), 1489–1506 (2000)
- [4] Sebe, N., Tian, Q., Loupias, E., et al.: Evaluation of salient point techniques. *Image and Vision Computing* 21(13-14), 1087–1095 (2003)
- [5] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
- [6] Qiu, B., Racoceanu, D., Xu, C.S., Tian, Q.: Stripe: image feature based on a new grid method and its application in ImageCLEF. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) *AIRS 2006. LNCS*, vol. 4182, Springer, Heidelberg (2006)
- [7] Ranawana, R., Palade, V.: Multi classifier systems – a review and roadmap for developers. *Journal of Hybrid Intelligent Systems* (2006)
- [8] Müller, H., et al.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: *Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, September 20-22 (2006)

Inter-media Concept-Based Medical Image Indexing and Retrieval with UMLS at IPAL

Caroline Lacoste, Jean-Pierre Chevallet, Joo-Hwee Lim, Diem Thi Hoang Le, Wei Xiong, Daniel Racoceanu, Roxana Teodorescu, and Nicolas Vuillenemot

IPAL International Joint Lab
Institute for Infocomm Research (I2R)
Centre National de la Recherche Scientifique (CNRS)
21 Heng Mui Keng Terrace
Singapore 119613

{viscl,viscjp,joothwee,stuthdl,wxiong,visdaniel}@i2r.a-star.edu.sg

Abstract. We promote the use of explicit medical knowledge to solve retrieval of information both visual and textual. For text, this knowledge is a set of concepts from a Meta-thesaurus, the Unified Medical Language System (UMLS). For images, this knowledge is a set of semantic features that are learned from examples using SVM within a structured learning framework. Image and text index are represented in the same way: a vector of concepts. The use of concepts allows the expression of a common index form: an inter-media index, offering the opportunity of homogeneous indexing/querying time fusion techniques. Top results obtained with concept based approaches show the potential of conceptual indexing.

1 Introduction

Medical Content-Based Image Retrieval (CBIR) systems has the potential to assist doctors in diagnosis by retrieving images with known pathologies that are similar to a patient's images. A Medical CBIR is also useful for teaching and research. Much existing CBIR systems [1] use primitive visual features to index visual content such as color or texture. We promote the use of explicit knowledge, to enhance the usual poor results of CBIR in the medical domain like specialized systems [2].

Pathology-bearing regions tend to be highly localized. Hence, local features such as those extracted from segmented dominant image regions seem an efficient solution. However, it has been recognized that pathology bearing regions are difficult to segment out automatically for many medical domains [2]. Hence we believe is desirable to have a medical CBIR system that does not rely on robust region segmentation and indexes images from learned image samples.

On the textual side, combining text and image generally increases retrieval performance. Simple approach consists in combining several Retrieval Engines at retrieval time. Fusion of retrieval results has been largely studied mainly for text

engine [3] and usually at querying time. In this paper, we present our work on medical image retrieval that is mainly based on the incorporation of medical knowledge in the system. For text, this knowledge is the one contained the Unified Medical Language System (UMLS) Meta-Thesaurus produced by NLM¹. For images, this knowledge is in semantic features that are learned from samples and do not rely on robust region segmentation. We experimented with two complementary visual indexing approaches: a global indexing to access image modality, and a local indexing to access semantic local features. This local indexing does not rely on region segmentation but builds upon patch-based semantic detector [4].

We propose to represent both image and text in a common inter-media representation. In our system, this inter-media index is a weighed vector of UMLS concepts. The use of UMLS concepts allows our system some degree of abstraction (e.g. language independent) but also unify the way index are represented. To build this common index representation, fusion is necessary, and can be done either at indexing time, fusion of indexes, or at querying time, fusion of weighted document retrieved lists. We have experimented both approaches with a visual modality filtering, designed to remove visually aberrant images according to the query modality concept.

2 Concept-Based Textual Indexing

Concept based textual indexing consists on selecting concepts from a concept set relevant to document to be indexed. In order to build a conceptual index, we need:

- Terminology: is a list of terms (single or multi term) from a given domain in a given language. Terms are coming from actual language usage in the domain. They are generally stable noun phrases (i.e. less linguistic variations than any other noun phrase), and they should have an unambiguous meaning in the restricted domain they are used;
- Vocabulary: is a set of concepts. In our context, a concept is a language independent meaning (with a definition), associated with at least one term of the terminology.
- Conceptual Structure: each term is associated to at least one concept. Concepts are also organized into several networks. Each network links concepts using a conceptual relation;
- Conceptual Mapping Algorithm: it selects a set of potential concepts from a sentence using the Terminology and the Conceptual Structure.

We have chosen UMLS as a terminological and conceptual data source to support our indexing. It includes more 5.5 millions of terms in 17 languages and 1.1 millions of unique concepts. UMLS is a “meta thesaurus”, i.e. a merge of different sources (thesaurus, terms lists), and is neither complete, nor consistent. In particular, links among concepts are not equally distributed. It is not an ontology,

¹ National Library of Medicine -<http://www.nlm.nih.gov/>

because there is no formal description of concepts. Nevertheless this large set of terms from medical domain enables us to experiment a full scale conceptual indexing system. In UMLS, all concepts are assigned to at least one semantic type from the Semantic Network. This provides consistent categorization of all concepts in the meta-thesaurus at a relatively general level.

Despite the large set of terms and term variations available in UMLS, it still cannot cover all possible (potentially infinite) term variations. So for English texts, we use MetaMap[5] provided by NLM that acts as the conceptual mapping algorithm. We have developed a similar simplified tool for French and German documents. Take notice these concept extraction tools do not provide any disambiguation, also concepts extraction is limited to noun phrase: i.e. verbs are not treated.

Extracted concepts are then organized in conceptual vectors, like a conventional Vector Space Model (VSM). We use VSM weighting scheme provided by our XIOTA indexing system [6]. We have chosen the classical $tf \cdot idf$ measure for concept weight, because we made the assumption concepts statistical distribution to be similar to terms distribution. As text indexing is a separate process for the three languages, concept vectors can either be fused, or queried separately: in that case, separate query results are fused. We have chosen this second solution.

Table 1. Results of textual runs

Rank	run ID	MAP	R-prec
1/31	Textual_CDW	0.2646	0.3093
2/31	Textual_CPRF	0.2294	0.2843
3/31	Textual_CDF	0.2270	0.2904
5/31	Textual_TDF	0.2088	0.2405
10/31	Textual_CDE	0.1856	0.2503

Table 2. Results of automatic visual runs

Rank	run ID	MAP	R-prec
3/11	Visual_SPC+MC	0.0641	0.1069
4/11	Visual_MC	0.0566	0.0912
6/11	Visual_SPC	0.0484	0.0847

One major criticism we have against VSM is the lack of structure of the query. It seems obvious to us that it is the *complete* query that should be solved and not only part of it. After query examination, we found out that queries are implicitly structured according to some semantic types (e.g. anatomy, pathology, modality). We call this the "semantic dimensions" of the query and manage them using an extra "semantic dimension filtering" step. This filtering retains answers that incorporate at least one dimension. We use semantic structure on concepts provided by UMLS. Semantic dimension of a concept is then defined by its UMLS semantic type, grouped into semantic groups: Anatomy, Pathology and Modality. This correspond to the run "Textual_CDF" of Table 1. We also tested a similar dimension filtering based MESH terms (run "Textual_TDF"). In this case, the association between MESH terms and a dimension had to be done manually. Using UMLS concepts and an automatic concept type filtering improves the results.

We also tested query semantic structure by re-weighting answers according to dimensions (DW). Here, Relevance Status Value (RSV) output from VSM is multiplied by the number of concepts matched with the query according to the dimensions. This re-weighting scheme strongly emphasizes the presence of maximum number of concepts related to semantic dimensions and implicitly do the previous dimension filtering (DF)². This produces our best results of 2006 ImageCLEFmed with 0.26 of Mean Average Precision (MAP) and outperforms any other classical textual indexing reported in ImageCLEFmed 2006. Hence we have shown here the potential of conceptual indexing.

In run “Textual_CPRF”, we test Pseudo-Relevance Feedback (PRF). From the result of late fusion of text-image retrieval results, three top relevant documents retrieved are taken and all concepts of these documents are added into query for query expansion. Then, a dimension filtering is applied. In fact, this run should have been classified in the mixed runs as we also use the image information to have a better precision in the three first images. We also tested document expansion using the UMLS semantic network. Based on UMLS hierarchical relationships, each database concept is expanded by concepts positioned at a higher level in the UMLS hierarchy and connected to this concept with respect to the semantic relation “is a”. The expanded concepts have a higher position than document concept in UMLS hierarchy. For example a document indexed by the concept “molar teeth” would be also indexed by the more general concept “teeth”. This document would be thus retrieved if a user asks for a teeth photography. This expansion is not effective (“Textual_CDE”) is 4 points below dimension filtering.

3 Concept-Based Visual Indexing

We developed a structured learning framework to index images using visual medical (VisMed) terms [7]. They are typical patches characterized by a visual appearance in medical image regions and having an unambiguous meaning. Moreover VisMed term is expressed in the inter-media space as a combination of UMLS concepts. In this way, we have a common inter-media conceptual indexing language. We developed two complementary indexing approaches:

- a global conceptual indexing to access image modality: chest X-ray, gross photography of an organ, microscopy, etc.;
- a local conceptual indexing to access local features that are related to modality, anatomy, and pathology concepts.

Global Conceptual Indexing. This global indexing is based on a two level hierarchical classifier according to mainly modality concepts. This modality classifier is learned from about 4000 images splitted in 32 classes: 22 grey level modalities, and 10 color modalities. Each indexing term is characterized by a modality, anatomy (e.g. chest X-ray, gross photography of an organ) and sometimes,

² Because the relevance value is multiplied by 0 when no corresponding dimension is found in documents.

a spatial concept (e.g. axial, frontal), or a color percept (color, grey). Training images come from CLEF database (about 2500 samples), from the IRMA³ database (about 300 samples), and from the web (about 1200 samples). Images from CLEF were selected after modality text concept extraction followed by a manual filtering. The first level of the classifier corresponds to a classification for grey level versus color images using the first three color moments in the HSV on the entire image. The second level corresponds to the classification of modality knowing the image is in the grey or the color cluster. For the grey level cluster, we use grey level histogram (32 bins), texture features (mean and variance of Gabor coefficients for 5 scales and 6 orientations), and thumbnails (grey values of 16x16 resized image). For the color cluster, we have adopted HSV histogram (125 bins), Gabor texture features, and thumbnails with zero-mean normalization. For each SVM classifier, we adopted a RBF kernel $\exp(-\gamma|x - y|^2)$ where $\gamma = \frac{1}{2\sigma^2}$ and with a modified city-block distance:

$$|x - y| = \frac{1}{F} \sum_{f=1}^F \frac{|x_f - y_f|}{N_f} . \quad (1)$$

where $x = \{x_1, \dots, x_F\}$ and $y = \{y_1, \dots, y_F\}$ are feature vectors; x_f, y_f are feature vectors of type f ; N_f is the feature vector dimension, and F is the number of feature types⁴. This just-in-time feature fusion within the kernel combines the contribution of color, texture, and spatial features equally [8]. Probability of a modality MOD_i for an image z is given by:

$$P(\text{MOD}_i|z) = \begin{cases} P(\text{MOD}_i|z, C)P(C|z) & \text{if } \text{MOD}_i \in C \\ P(\text{MOD}_i|z, G)P(G|z) & \text{if } \text{MOD}_i \in G \end{cases} \quad (2)$$

where C and G denote the color and the grey level clusters respectively, and the conditional probability $P(\text{MOD}_i|z, V)$ is given by:

$$P(c|z, V) = \frac{\exp^{\mathcal{D}_c(z)}}{\sum_{j \in V} \exp^{\mathcal{D}_j(z)}} . \quad (3)$$

where $\mathcal{D}_c$ is the signed distance to the SVM hyperplane that separate class c from the other classes of the cluster V .

After learning using SVM-Light [9], each image z is indexed according to modality given its low-level features z_f . Indexes weight are probability values given by Equation (2).

Local Conceptual Indexing. Local indexing uses Local Visual Patches (LVP). LVP grouped in Local Visual Concepts (LVC) linked with concepts from Modality, Anatomy, and Pathology semantic types. The patch classifier uses 64 LVC.

³ http://phobos.imib.rwth-aachen.de/irma/index_en.php

⁴ $F = 1$ for the grey versus color classifier, $F = 3$ for the conditional modality classifiers: color, texture, thumbnails.

In these experiments, we have adopted color and texture features, SVM classifier with same parameters as global indexing. Color features are the three first color moments of Hue, Saturation, and Value. Texture is mean and variance of Gabor coefficients using 5 scales and 6 orientations. Zero-mean normalization is applied. The training dataset is composed of 3631 LVP extracted from 1033 images mostly coming from the web (921 images coming from the web and 112 images from the CLEF collection $\sim 0.2\%$).

After learning, LVC are identified during image indexing using image patches without region segmentation to form a LVC histogram. Essentially, an image is tessellated into overlapping image blocks of size 40x40 pixels after size standardization [4]. Each patch is then classified into one of the 64 LVC using the Semantic Patch Classifier. Histogram aggregation per block gives the final image index. Each bin of the LVC histogram of a given block B corresponds to the probability of a LVC_i presence in this block. This probability is computed as follows:

$$P(LVC_i|B) = \frac{\sum_z |z \cap B| P(LVC_i|z)}{\sum_z |z \cap B|} . \quad (4)$$

where B is an image block, z an image patch, $|z \cap B|$ the intersection area between z and B , and $P(LVC_i|z)$ is given by Equation (3).

Visual Conceptual Retrieval. We propose three retrieval methods based on the two indexing. When several images are given in the query, fuzzed similarity is the maximum of each similarity values. The first method (“Visual_MC”) corresponds to the global indexing scheme. An image is represented by a concept histogram, each bin corresponding to a modality probability. We use Manhattan distance between two histograms. The second method (“Visual_SPC”) uses the local UMLS visual indexing. Distance is the mean of block by block Manhattan distances on all the possible matches. The last visual retrieval method (“Visual_SPC+MC”) is the mean fusion of the two first methodes: global and local.

The 2006 CLEF medical task is particularly difficult: the best automatic visual result was less than 0.08 of MAP. Mixing the local and global indexing, is our best solution with 0.06 of MAP⁵ as showed in Table 2.

Manual query construction gives by far better results (between 0.1596 and 0.1417 of MAP), but can still be consider as low regarding the human effort and implication into the system to set up queries and produce relevance feedback. See [10] for more details.

4 Inter-media Retrieval and Fusion

We propose three types of fusion between conceptual based visual and textual indexes: querying time linear and visual filtered fusion using modality concept,

⁵ The Mean Average Precision and the R-precision computed in imageCLEFmed for the run Visual_SPC+MC was 0.0634 and 0.1048 respectively because we only submitted by error the 25 first queries.

and indexing time fusion. The querying time fusion combines visual and textual similarity measures: the Relevance Status Value (RSV) between a mixed⁶ query $Q = (Q_I, Q_T)$ and mixed document $D = (D_I, D_T)$ is then computed by an normalized weighted linear combination of RSV:

$$RSV(Q, D) = \alpha \frac{RSV_I(Q_I, D_I)}{\max_{z \in \mathcal{D}_I} RSV_I(Q_I, z)} + (1 - \alpha) \frac{RSV_T(Q_T, D_T)}{\max_{z \in \mathcal{D}_T} RSV_T(Q_T, z)} \quad (5)$$

where RSV_V is the maximum of visual similarity between all images of Q_I and all images of D_I , RSV_T is the textual Relevance Status Value. $\mathcal{D}_I$ denotes the image database, and $\mathcal{D}_T$ denotes the text database. The factor α allows control of the weight between textual image similarity. After some experimentations on imageCLEFmed 2005, we chose $\alpha = 0.7$. The result “Cpt_Im” in Table 3 is the overall best combination. It shows the good complement of the visual and textual indexing: from 0.26 for the textual retrieval and 0.06 for the visual retrieval, the mixed retrieval provides 0.31 of MAP. The second type of

Table 3. Results of automatic mixed runs

Rank	run ID	MAP	R-prec
1/37	Cpt_Im	0.3095	0.3459
2/37	ModFDT_Cpt_Im	0.2878	0.3352
3/37	ModFST_Cpt_Im	0.2845	0.3317
4/37	ModFDT_TDF_Im	0.2730	0.3774
5/37	ModFDT_Cpt	0.2722	0.3757
17/37	MediSmart_1	0.0649	0.1012
30/37	MediSmart_2	0.0420	0.0657

fusion implies a common inter-media conceptual indexing for filtering: both index (image and text) are in the same Conceptual Vector Space. A comparison between query concepts related to modality and image modality index is done in order to remove all aberrant images using a threshold. We tested this approach with, first, a fixed threshold for all modality (“ModFST”), and, second, an adaptive threshold for each modality according a confidence degree given to the classifier according this modality (“ModFDT”). These two methods increase the results of the purely textual retrieval. Finally, indexing time fusion has been experimented with a fuzzyfication algorithm that takes into account concept frequency, document parts localization and confidence of concept extraction (see [11] for more details). Official submission was erroneous “MediSmart”: we now obtain 0.2568 of MAP without dimension filtering. This good result demonstrates the potentiality of Inter-Media Conceptual Indexing with an indexing time fusion.

⁶ I for image and T for text.

5 Conclusion

In this paper, we have proposed a medical image retrieval system that represents both texts and images in the same conceptual inter-media space. This enables the filtering and fusion at this level, and enables an early fusion at indexing time. With top results, we demonstrate the efficiency of the use of explicit knowledge and conceptual indexed for text. This solution is less effective on images. Even if currently our best results are obtained using querying time fusion, we believe that conceptual inter-media indexing time fusion is a new approach with a strong potential that need to be investigated.

This work is part of the ISERE ICT Asia project, supported by the French Minister of Foreign Affairs.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1349–1380 (2000)
2. Shyu, C.R., Pavlopoulou, C., Kak, A.C., Brodley, C.E., Broderick, L.S.: Using human perceptual categories for content-based retrieval from a medical image database. *Computer Vision and Image Understanding* 88(3), 119–151 (2002)
3. Montague, M., Aslam, J.A.: Condorcet fusion for improved retrieval. In: *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pp. 538–548. ACM Press, New York (2002)
4. Lim, J., Chevallet, J.P.: Vismed: a visual vocabulary approach for medical image indexing and retrieval. In: Lee, G.G., Yamada, A., Meng, H., Myaeng, S.-H. (eds.) *AIRS 2005. LNCS*, vol. 3689, pp. 84–96. Springer, Heidelberg (2005)
5. Aronson, A.: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: *Proceedings of the Annual Symposium of the American Society for Medical Informatics*, pp. 17–21 (2001)
6. Chevallet, J.P.: X-IOTA: An open XML framework for IR experimentation application on multiple weighting scheme tests in a bilingual corpus. In: Myaeng, S.-H., Zhou, M., Wong, K.-F., Zhang, H.-J. (eds.) *AIRS 2004. LNCS*, vol. 3411, pp. 263–280. Springer, Heidelberg (2005)
7. Lim, J., Chevallet, J.: Vismed: a visual vocabulary approach for medical image indexing and retrieval. In: *Proceedings of the Asia Information Retrieval Symposium*, pp. 84–96 (2005)
8. Lim, J., Jin, J.: Discovering recurrent image semantics from class discrimination. *EURASIP Journal of Applied Signal Processing* 21, 1–11 (2006)
9. Joachims, T.: *Learning to Classify Text using Support Vector Machines*. Kluwer Academic Publishers, Dordrecht (2002)
10. Wei, X., Bo, Q., Qi, T., Changsheng, X., Sim-Heng, O., Kelvin, F.: Combining multilevel visual features for medical image retrieval in ImageCLEF 2005. In: *Cross Language Evaluation Forum 2005 workshop*, Vienna, Austria, September 2005, p. 73.
11. Racocceanu, D., Lacoste, C., Teodorescu, R., Vuilleminot, N.: A semantic fusion approach between medical images and reports using UMLS. In: *Proceedings of the Asia Information Retrieval Symposium (Special Session)*, Singapore (2006)

UB at ImageCLEFmed 2006

Miguel E. Ruiz

State University of New York at Buffalo
Department of Library and Information Studies
534 Baldy Hall
Buffalo, NY 14260-1020 USA
meruiz@buffalo.edu,
<http://www.informatics.buffalo.edu/faculty/ruiz>

Abstract. This paper presents the results of the University at Buffalo in the 2006 ImageCLEFmed task. Our approach for this task combines Content Based Image Retrieval (CBIR) and text retrieval to improve retrieval of medical images. Our results are comparable to other approaches presented in the task. Our results show that text retrieval performs well across the three different types of topics (visual, visual-semantic and semantic) and that the combination of CBIR and text retrieval yields moderate improvements.

1 Introduction

This paper presents the results of the University at Buffalo, The State University of New York, in the ImageCLEFmed 2006 task. We used a similar approach to the one proposed in ImageCLEFmed 2005 [2, 3]. This approach combines Content-Based Image Retrieval (CBIR) and text-retrieval by using a fusion formula of results retrieved by both approaches. Sections 2 and 3 present the preprocessing of the document and query topics. Section 4 describes the retrieval model used and tuning of the system parameters. Section 5 presents our results and analysis. The last section presents our conclusions and future work.

2 Collection Preparation

The text associated with the images was parsed and converted into our SMART format. We indexed all the information available in the text descriptions into a single text field. Image ids associated to this text were also added in a separate ctype.

The text data was indexed using the SMART retrieval system [5], which was modified in house to support ISO-Latin-1 encoding and the newest weighting schemes such as pivoted length normalization and BM25. This system has been used in previous CLEF participations of the UB team [3, 4]. We created a single index that includes all English, French, and German text. A single list of all tree languages was used to discard stop-words. The remaining terms were processed with a basic stemmer that normalizes words in plural form (Remove-S).

Images were indexed using GIFT standard settings to process images with 4 grey levels and the full HSV space using the Gabor filter responses in four directions and at three scales.

3 Query Processing

Queries were indexed using a similar approach to the one used for documents: discarding stop words and stemming plurals. For this year we used the version of the queries that use all three languages. In other words, we did not do any automatic query translation in this runs.

4 Retrieval Model

We performed automatic retrieval feedback by retrieving 1000 documents using the original query and assuming that the top n documents are relevant and the bottom 100 documents are not relevant. This allowed us to select the top m terms ranked according to Rocchio's relevance feedback formula [1]:

$$w_{new}(t) = \alpha \times w_{orig}(t) + \beta \times \frac{\sum_{i \in Rel} w(t, d_i)}{\|Rel\|} - \gamma \times \frac{\sum_{i \in \neg Rel} w(t, d_i)}{\|\neg Rel\|} \quad (1)$$

where α , β , and γ are coefficients that control the contribution of the original query, the relevant documents (Rel) and the non-relevant documents ($\neg Rel$) respectively. The optimal values for these parameters are also determined using the CLEF 2005 topics.

4.1 Combining Visual and Text Retrieval

The system submits the image(s) associated to the topic to GIFT and the text to the SMART system. The final step consists in combining the list of images retrieved by the GIFT with the images associated to the case descriptions retrieved by SMART. We use the following formula to compute the final score of each image:

$$w_k = \lambda \times I_k + \mu \times T_k \quad (2)$$

where I_k is the score of the image k in the CBIR system and T_k is the score of the image k in the text retrieval system. λ and μ are parameters that control the contribution of each system to the final score. Note that this formula assumes that the scores from both systems are normalized between 0 and 1. In our experiments we used a tf-idf schema for GIFT and SMART which guaranties that the scores are between 0 and 1.

The parameters for the retrieval model were adjusted using the data available for ImageCLEFmed 2005. The optimal parameters obtained were $\alpha = 8$, $\beta = 64$ and $\gamma = 16$ for the Rocchio's formula and $\lambda = 1$ and $\mu = 3$ for the weights associated to the image and text scores respectively.

5
Results

Our results are presented in Table 1. We submitted 4 official runs: 2 baseline runs that use textual features only and 2 runs that use the combined results of the CBIR and Text retrieval system. Our results are above the median runs in this track. In general the Text retrieval system performs well above our CBIR baseline. However, the combination of text and visual runs shows modest improvements. Table 2 and Figure 1 show the performance of the CBIR, text-only,

Table 1. Performance of official runs

Run Label	Mean Avg. Precision Parameters	
UBmedTBL1 (text only)	0.1894	$n = 10, m = 10$
UBmedT1 (text only)	0.1965	$n = 50, m = 10$
UBmedVT1 (Visual and Text)	0.1938	$n = 10, m = 10$ $\lambda = 1, \mu = 3$
UBmedVT2 (Visual and Text)	0.2027	$n = 50, m = 10$ $\lambda = 1, \mu = 3$

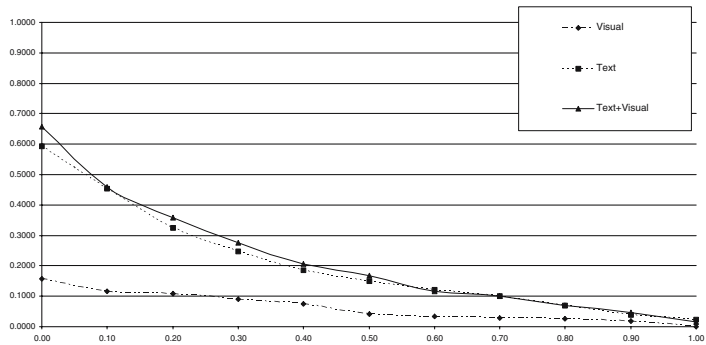


Fig. 1. Comparison of CBIR (Visual), Text and Combined Text & Visual runs

Table 2. Comparison of performance by type of topic

Query Type	CBIR (Visual)	SMART (text-only)	Combined (Visual + Text)
Visual	0.0596	0.1619	0.1321
V-S	0.0393	0.1882	0.2241
Semantic	0.0348	0.2314	0.2521
All	0.0446	0.1938	0.2027

and the combined text-visual runs. These show that the combined runs improve results for visual-semantic and semantic topics.

6 Conclusion

Our results confirm that combining CBIR and text retrieval improves average performance when compared to using either text only or CBIR. Although the improvement obtained is relatively modest it is interesting to note that for some types of queries such as the visually-semantically possible this improvement is relatively large.

Our future plans include exploring other aspects that could yield larger improvements in performance such as image classification and topic characterization before retrieval.

References

- [1] Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Englewood Cliff, NJ (1971)
- [2] Ruiz, M.: Combining image features, case descriptions and umls concepts to improve retrieval of medical images. In: *Proceedings of the American Medical Informatics Association 2006 Annual Symposium*, Washington, DC, pp. 674–678 (2006)
- [3] Ruiz, M., Southwick, S.: Ub at clef 2005: Bilingual portuguese and medical image retrieval tasks. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
- [4] Ruiz, M., Srikanth, M.: Ub at clef 2004: Cross language medical image retrieval. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004. LNCS*, vol. 3491, pp. 773–780. Springer, Heidelberg (2005)
- [5] Salton, G. (ed.): *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, NJ (1983)

Translation by Text Categorisation: Medical Image Retrieval in ImageCLEFmed 2006

Julien Gobeill, Henning Müller, and Patrick Ruch

Medical Informatics, University and Hospitals of Geneva, Switzerland
{julien.gobeill, henning.mueller, patrick.ruch}@sim.hcuge.ch

Abstract. We present the fusion of simple retrieval strategies with thesaural resources to perform document and query translation by text categorisation for cross-language retrieval in a collection of medical images with case notes. The collection includes documents in French, English and German. The fusion of visual and textual content is also treated. Unlike most automatic categorisation systems our approach can be applied with any controlled vocabulary and does not require training data. For the experiments we use Medical Subject Headings (MeSH), a terminology maintained by the National Library of Medicine existing in 12 languages. The idea is to annotate every text of the collection (documents and queries) with a set of MeSH terms using our automatic text categoriser. Our results confirm that such an approach is competitive. Simple linear approaches were used to combine text and visual features.

1 Introduction

Cross-Language Information Retrieval (CLIR) is increasingly relevant as network-based resources become commonplace. In the medical domain it is of strategic importance to fill the gap between clinical records in national languages and research reports in English. Images are also getting increasingly important and varied in the medical domain, and they become available in digital form. Despite images being language-independent, they are most often accompanied by textual notes in various languages and these notes can improve retrieval quality [1]. A task description on the medical image retrieval task can be found in [2].

2 Strategies

1. Each document and query are annotated by our automatic text categoriser, which contains MeSH in French, English, and German.
2. Each query is annotated by 3 terms and each document by 3, 5, 8 categories.

2.1 Terminology-Driven Text Categorisation

Automatic text categorisation has been studied largely and led to a large amount of papers. Approaches include naive Bayes, k-nearest neighbours, boosting, rule-learning algorithms. However, most of these studies apply text classification to

a small set of classes. In comparison to this our system is designed to handle large class sets [3], since such retrieval tools used are only limited by the size of the inverted file, but 10^{5-6} documents is a modest range.

Our approach is data-poor as it requires only a small collection of annotated texts for tuning: instead of inducing a complex model using training data, our categoriser indexes the collection of MeSH (Medical Subject Headings) terms as if they were documents and then it treats the input as if it was a query to be ranked regarding each MeSH term. The classifier is tuned by using English abstracts. For the tuning, the top 15 returned terms are selected because it is the average number of MeSH terms per abstract in the OHSUMED collection.

2.2 Regular Expressions and MeSH Thesaurus

The regular expression search tool is applied on the canonical MeSH collection augmented with a list of MeSH synonyms. In this system, string normalisation is mainly performed by the MeSH terminological resources when the thesaurus is used. The MeSH provides a large set of related terms, which are mapped to a unique MeSH representative in the canonic collection. The related terms gather morpho-syntactic variants, strict synonyms, and a last class of related terms, which mixes up generic and specific terms: for example, *Inhibition* is mapped to *Inhibition (Psychology)*. The system cuts the abstract into 5-token-long phrases and moves the window through the abstract: the edit-distance is computed between each of these 5-token sequences and each MeSH term. Basically, the manually crafted finite-state automata allow two insertions or one deletion within a MeSH term, and ranks the proposed candidate terms based on these basic edit operations: insertion costs 1, while deletion costs 2. The resulting pattern matcher behaves like a term proximity scoring system [4], but is restricted to a 5-token matching window.

2.3 Vector-Space Classifier

The vector-space (VS) module is based on a general IR engine with the *tf.idf* weighting schema. The engine uses a list of 544 stop words. For setting the weighting factors, we observed that cosine normalisation was effective for our task. This is not surprising because cosine normalisation performs well when documents have a similar length [5]. For the respective performance of each basic classifier, the RegEx system performs better than any *tf.idf* schema, so the pattern matcher provides better results than the VS engine. However, we also observe that the VS system gives better precision at high ranks (*Precision_{at Recall=0}* or *mean reciprocal rank*) than the RegEx system: this difference suggests that merging the classifiers could be effective.

2.4 Classifier Fusion

The hybrid system combines the regular expression classifier with the VS classifier. Unlike [6] we do not merge our classifiers by linear combination, because

the RegEx module does not return a scoring consistent with the VS system. Therefore, the combination does not use the RegEx's edit distance, and instead uses the list returned by the VS module as a *reference* list (*RL*), while the list returned by the regular expression module is used as *boosting* list (*BL*), which serves to improve the ranking of terms listed in *RL*. A third factor takes into account the term length: both the number of characters (L_1) and the number of tokens (L_2 , with $L_2 > 3$) are computed, so that long and compound terms, which appear in both lists, are favoured over single and short terms. For each concept t listed in the *RL*, the combined Retrieval Status Value (*cRSV*, equation 1) is:

$$cRSV_t = \begin{cases} RSV_{VS}(t) \cdot \ln(L_1(t) \cdot L_2(t) \cdot k) & \text{if } t \in BL, \\ RSV_{VS}(t) & \text{otherwise.} \end{cases} \quad (1)$$

The value of the k parameter is set empirically.

2.5 Cross-Language Categorisation and Indexing

To translate the ImageCLEFmed text, we use the English MeSH categorisation tool. French, and German versions of the MeSH are simply merged in the categoriser. We use the weighting schema combination (ltc.lnn + RegEx). Then, the annotated collection is indexed using the VS engine. For document indexing, we rely on weighting schemas based on pivoted normalisation: as documents have a very variable length such a factor can be important. A slightly modified version of dtu.dtn [7] is used for full-text indexing and retrieval. The English stop word list is merged with a French and a German stop word list. Porter stemming is used for all documents.

2.6 Visual and Multimodal Retrieval

Visual retrieval is mainly based on *GIFT*¹ [8]. Features used by *GIFT* are:

- Local color features at various scales by partitioning the images successively into four equally sized regions and taking the mode color of each region;
- global color features in the form of a color histogram, compared by a histogram intersection;
- local texture features by partitioning the image and applying Gabor filters in various scales and directions, quantised into 10 strengths;
- global texture features represented as a simple histogram of responses of the local Gabor filters in various directions and scales.

The feature space is similar to the distribution of words in texts. A *tf/idf* weighting is used and the query weights are normalised by the results of the query.

To combine visual and textual runs we choose English as language and a number of five terms based on visual observations. The combination is done by normalising the output of visual and textual results and adding them up in various ratios. A second approach for a multimodal combination was to take results from the visual retrieval side and increase the value of those results in the first 1000 images that also appear in the visual results.

¹ <http://www.gnu.org/software/gift/>

3 Results and Discussion

3.1 Textual Retrieval Results

The runs were generated using 8, 5, and 3 MeSH terms to annotate the collection. In previous experiments [9], the optimal number was around 2 or 3.

Table 1. MAP of textual runs

Run	MAP	Run	MAP	Run	MAP
GE-8EN	0.2255	GE-8DE	0.0574	GE-8FR	0.0417
GE-5EN	0.1967	GE-5DE	0.0416	GE-5FR	0.0346
GE-3EN	0.1913	GE-3DE	0.0433	GE-6FR	0.0323

Results are computed by retrieving 1'000 documents per query. In Table 1, we observe that the maximum MAP is reached when eight MeSH terms are selected per document. This suggests that a larger number can be selected to annotate a document, although it must be observed that the precision of the system is low beyond the top ranked (one or two) categories. This means that annotating a document with several potentially irrelevant concepts does not hurt the matching power! This result is consistent with known observations made on query expansion: some inappropriate expansion is acceptable and can still improve retrieval effectiveness. The English retrieval results were the second best results of all participants². For other languages it seems to be much harder to obtain good results as the majority of documents is in English.

3.2 Visual and Multi-modal Runs

Table 2 shows the results for our visual run and the best mixed runs. The visual run is performance-wise in the middle of the submissions and the best purely visual runs are approximately 30% better. GIFT performs better in early precision than other systems with a higher MAP. For the visual topics the results are very satisfactory whereas semantic topics do not perform well.

Table 2. MAP of visual runs

Run	MAP
GE-GIFT	0.0467
GE-vt10	0.12
GE-vt20	0.1097

A problem shows up in all mixed runs submitted. They are worse than the underlying textual runs even when only a small fraction of visual information is used. A possible problem is the use of a wrong file for the text runs. English runs perform much better than French and German runs. We need to further investigate to find the reasons and allow for better multimodal image retrieval.

² Some groups combined results of the three languages as it was not forbidden, which is not a realistic scenario although it improves results.

4 Conclusion and Future Work

We presented a cross-language information retrieval engine for the ImageCLEFmed retrieval task, which uses a multilingual controlled vocabulary to translate user requests and documents. The system relies on a text categoriser, which maps queries into a set of predefined concepts. For ImageCLEFmed 2006 optimal precision is obtained when selecting three MeSH terms per query and eight per document, whereas a larger number might even improve results further. Visual retrieval shows to work well on visual topics but fails on semantic topics. A problem is the combination of visual and textual features that needs further analysis and best an analysis of the query to find out more about the search goal.

Acknowledgements

The study was supported by the Swiss National Foundation (3200-065228, 205321-109304/1) and the EU (SemanticMining NoE, INFS-CT-2004-507505).

References

1. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine – clinical benefits and future directions. *International Journal of Medical Informatics* 73, 1–23 (2004)
2. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Eugene, K., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In: *CLEF 2006 Proceedings. LNCS* (to appear, 2007)
3. Ruch, P.: Automatic Assignment of Biomedical Categories: Toward a Generic Approach. *Bioinformatics* 6, 658–664 (2006)
4. Rasolofo, Y., Savoy, J.: Term proximity scoring for keyword-based retrieval systems. In: *European Conference on Information Retrieval*, pp. 101–116 (2003)
5. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: *Proceedings of the ACM SIGIR Conference*, pp. 21–29. ACM Press, New York (1996)
6. Larkey, L., Croft, W.: Combining classifiers in text categorization. In: *Proceedings of the ACM SIGIR Conference*, pp. 289–297. ACM Press, New York (1996)
7. Aronson, A., Demner-Fushman, D., Humphrey, S., Lin, J., Liu, H., Ruch, P., Ruiz, M., Smith, L., Tanabe, L., Wilbur, J.: Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. *Proceedings of TREC 2005* (2006)
8. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters* 21, 1193–1198 (2000)
9. Ruch, P.: Query translation by text categorization. In: Kaufmann, A.M. (ed.) *Proceedings of COLING 2004*, pp. 686–692 (2004)

Using Information Gain to Improve the ImageCLEF 2006 Collection

M.C. Díaz-Galiano, M.Á. García-Cumbreras, M.T. Martín-Valdivia,
A. Montejo-Ráez, and L. Alfonso Ureña-López

SINAI Research Group. Computer Science Department. University of Jaén. Spain
`{mcdiaz,magc,maite,amontejo,laurena}@ujaen.es`

Abstract. This paper describes the SINAI team's participation in both the ad hoc task and the medical task. For the ad hoc task we use a new Machine Translation system which works with several translators and heuristics. For the medical task, we have processed the set of collections using Information Gain (IG) to identify the best tags that should be considered in the indexing process¹.

1 The Ad Hoc Task

The goal of the ad hoc task is, given a multilingual query, to find as many relevant images as possible from a given image collection [2].

1.1 Description of Experiments

We have considered seven languages in our experiments: Dutch, English, French, German, Italian, Portuguese and Spanish. Based on our reliable 2005 results [3], this year we have used the same IR system and the same strategies, except that we used a new translation module. This module combines the following Machine Translators and heuristics:

- Epals (German and Portuguese), Prompt (Spanish), Reverso (French) and Systran (Dutch and Italian)
- Some heuristics are, for instance, the use of the translation made by the translator by default, a combination with the translations of every translator, or a combination of the words with a higher punctuation (scoring two points if it appears in the default translation, and one point if it appears in all of the other translations).

The collections have been preprocessed using stopwords and the Porter's stemmer. The collection dataset has been indexed using the LEMUR IR system², using two different weighting schemes (Okapi or TFIDF), and enabling or disabling PRF (pseudo-relevance feedback).

¹ This work has been supported by the Spanish Government with grant TIC2003-07158-C04-04.

² <http://www.lemurproject.org/>

1.2 Results and Discussion

All the results are obtained using the title and narrative text, whenever possible. In the English monolingual task and in the German-English bilingual task, we have combined the use (or not) of pseudo-relevance feedback and the weighting function (Okapi or Tfidf). In table 1, we can see the global results: English monolingual results and bilingual results. They show that the pseudo-relevance feedback is important when Okapi is used as a weighing function. The results with Tfidf and with Okapi without PRF are very poor.

The results also show that there is a loss of MAP between the best monolingual experiment and this bilingual experiment, namely approximately 28%. Yet, the other results in the English monolingual task are less acceptable compared to the German bilingual ones.

In general, there is a loss of precision compared to the English monolingual results. The Spanish result is approximately 17% less acceptable. The other languages lead to a deterioration of the results.

Table 1. Summary of results for the English monolingual adhoc runs and the other bilingual runs

Language	Initial Query	Expansion	Weight	MAP	Rank
Monolingual En	title + narr	with	Okapi	0.2234	9/49
Monolingual En	title + narr	without	Okapi	0.0845	38/49
Monolingual En	title + narr	with	Tfidf	0.0846	37/49
Monolingual En	title + narr	without	Tfidf	0.0823	39/49
Bilingual DeEn	title + narr	with	Okapi	0.1602	4/8
Bilingual DeEn	title + narr	without	Okapi	0.1359	7/8
Bilingual DeEn	title + narr	with	Tfidf	0.1489	5/8
Bilingual DeEn	title + narr	without	Tfidf	0.1369	6/8
Bilingual NlEn(Dutch)	title + narr	with	Okapi	0.1261	4/4
Bilingual FrEn(French)	title + narr	with	Okapi	0.1617	5/8
Bilingual ItEn(Italian)	title + narr	with	Okapi	0.1216	13/15
Bilingual PtEn(Portuguese)	title + narr	with	Okapi	0.0728	7/7
Bilingual EsEn(Spanish)	title + narr	with	Okapi	0.1849	4/7

2 The Medical Task

This year our experiments focus on preprocessing the collection using Information Gain (IG) in order to improve the quality of results and to automate the tag selection process.

In order to generate the textual corpus we have preprocessed the collection using the ImageCLEFmed.xml file [1]. The process is the same as last year [3].

However, this year the XML tags have been selected according to the amount of information theoretically supplied. For this reason, we have used the information gain measure to select the best tags in the collection.

The method applied consists in computing the information gain for every tag at every sub-collection. Since each subcollection has a different set of tags, the information gain was calculated for each subcollection individually. Let C be the set of cases and E the value set for the E tag, then the formula that we have to compute must obey the following expression:

$$IG(C|E) = H(C) - H(C|E) \quad (1)$$

where

$IG(C|E)$ is the information gain for the E tag,

$H(C)$ is the entropy and of the set of cases C

$H(C|E)$ is the relative entropy of the set of cases C conditioned by the E tag

Both, $H(C)$ and $H(C|E)$ are calculated based on the frequencies of occurrence of tags according to the combination of words which they represent. After some basic operations, the final equation for the computation of the information gain supplied by a given tag E over the set of cases C is defined as follows:

$$IG(C|E) = -\log_2 \frac{1}{|C|} + \sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|} \log_2 \frac{1}{|C_{e_j}|} \quad (2)$$

For every tag in every collection, its information gain is computed. Then, the tags selected to compose the final collection are those showing high values of IG. Once the document collection was generated, experiments were conducted with the LEMUR retrieval information system, applying the KL-divergence weighting scheme.

2.1 Experiment Description and Results

Our main goal is to investigate the effectiveness of filtering tags using IG in the text collection. To that end, we have accomplished several experiments using the ImageCLEFmed2005 to identify the best tag percentage with experiments preserving 10%, 20%...100% of tags. The results were evaluated with the relevance assessments of the 2005 collection. Only runs with 20%, 30% and 40% of tags for the 2006 collection were submitted, because these settings led to the best results on the 2005 corpus.

We also wanted to compare the results obtained when using only the text associated to the query topic with the results obtained when we merge visual and textual information. For this purpose, the first experiment was conducted as a baseline case. This experiment simply consists in taking the text associated to each query as a new textual query, which is then fed into the LEMUR system. The resulting list is the baseline run.

The remaining experiments start from the ranked lists provided by the GIFT tool for each query. For each list/query we have used an automatic textual query expansion using the associated text to the first four ranked images from the GIFT

lists and the original textual query in order to generate a new textual query. The new textual query is then submitted to the LEMUR system and a new ranked list is obtained. Finally, the expanded textual list and GIFT list are merged in order to obtain one final list (FL) with relevant images ranked by relevance. The merging process was done giving different importance to the visual (VL) and textual lists (TL): $FL = TL * \alpha + VL * (1 - \alpha)$

In order to adjust these parameters some experiments were accomplished with the 2005 collection varying α in the range [0,1] with step 0.1 (i.e., 0, 0.1, 0.2,...,0.9 and 1). After analyzing the results, we have submitted runs with α set to 0.5, 0.6 and 0.7 for the 2006 collection.

These three experiments and the baseline experiment (that only uses textual information of the query) have been accomplished over the three different corpora generated with 20%, 30% and 40% of tags. All textual experiments have been carried out with LEMUR using PRF and the Kl-divergence weighting scheme, as pointed out previously. Table 2 shows the results obtained with the SINAI system.

Table 2. Performance in Medical Image Retrieval

Runs	Experiment	Precision
	mixed SinaiGiftT50L20	0.0495
	mixed SinaiGiftT50L30	0.0491
	mixed SinaiGiftT50L40	0.0494
	mixed SinaiGiftT60L20	0.0468
	mixed SinaiGiftT60L30	0.0465
	mixed SinaiGiftT60L40	0.0470
	mixed SinaiGiftT70L20	0.0435
	mixed SinaiGiftT70L30	0.0437
	mixed SinaiGiftT70L40	0.0441
Runs	Experiment	Precision
text only	SinaiOnlytL20	0.1955
text only	SinaiOnlytL30	0.2167
text only	SinaiOnlytL40	0.2215

References

1. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the ImageCLEFmed, medical retrieval and annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS, Springer, Heidelberg (to appear, 2007)
2. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF, photographic retrieval and object annotation tasks. Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006 (to appear, 2006)
3. Martín-Valdivia, M.T., García-Cumbreras, M.Á., Díaz-Galiano, M.C., Ureña-López, L.A., Montejo-Ráez, A.: SINAI at ImageCLEF 2005. In: Proceedings of the Cross Language Evaluation Forum CLEF 2005 (2005)

CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections

M.M. Rahman¹, V. Sood¹, B.C. Desai¹ and P. Bhattacharya²

¹ Dept. of Computer Science & Software Engineering, Concordia University, Canada

² Concordia Institute for Information Systems Engineering, Concordia University, Canada*

Abstract. This paper presents the techniques and the analysis of different runs submitted by the CINDI group in ImageCLEF 2006. For the ad-hoc image retrieval from both the photographic and medical image collections, we experimented with cross-modal (image and text) interaction and integration approaches based on the relevance feedback in the form of textual query expansion and visual query point movement with adaptive similarity matching functions. For the automatic annotation tasks for both the medical and object collections, we experimented with a classifier combination approach, where several probabilistic multi-class support vector machine classifiers with features at different levels as inputs are combined to predict the final probability scores of each category as image annotation. Analysis of the results of the different runs we submitted for both the image retrieval and annotation tasks are reported in this paper.

Keywords: Multi-modality, Content-based image retrieval, Annotation, Relevance feedback, Fusion, Classifier combination.

1 Introduction

For the 2006 ImageCLEF workshop, CINDI research group has participated in four different tasks of ImageCLEF track: an ad-hoc retrieval from a photographic collection (e.g., IAPR dataset), ad-hoc retrieval from a medical (e.g., CASImage, MIR, PathoPic, and Peir datasets) collection, and automatic annotation of the medical (e.g., IRMA dataset) and the object (e.g., LTU dataset) collections [1,2]. This paper presents the methodologies, results and analysis of the runs of each of the tasks separately.

2 Ad-Hoc Retrieval in the Photographic and Medical Collections

Our main goal for the ad-hoc retrieval task is to investigate the effectiveness of combining keyword and content-based image retrieval (CBIR) approaches by

* This work was partially supported by NSERC, IDEAS and Canada Research Chair grants.

involving users in the retrieval loop in the form of relevance feedback (RF) [3]. We experimented with a cross-modal approach of image retrieval which integrates visual information based on purely low-level image content and semantical information from associated annotated text files. The advantages of both modalities are exploited by involving the users in the retrieval loop for the cross-modal interaction and integration in similarity matching.

For the CBIR approach, a query point movement and similarity matching function adjustment are performed based on the estimation of the means and covariance matrices from the feature vectors of the positive feedback images. For the text-based retrieval, the keywords from the annotated files are extracted and indexed with the help of the popular vector space model [4]. In order to perform the query expansion for textual search, additional keywords are appended for the query based on the relevant (positive) feedback from the user. Finally, an ordered list of images is obtained by using a pre-filtering approach which integrates the scores from the results for both the text and image-based search.

2.1 Content-Based Image Retrieval Approach

The performance of a CBIR system depends on the underlying image representation, usually in the form of a feature vector [5]. Based on the previous experiments [6], we found that the image features at different levels are complementary in nature and together they contribute effectively to distinguish the images of different visual and/or semantic categories. To generate the feature vectors, the low-level global, semi-global, low-resolution scaled and region specific local image features are extracted from the image representation at the different levels of abstraction.

In this work, the MPEG-7 based Edge Histogram Descriptor (EHD) and Color Layout Descriptor (CLD) [7] are extracted for the image representation at the global level. To represent the global shape feature, the spatial distribution of edges is utilized by the EHD descriptor. A histogram with $16 \times 5 = 80$ bins is obtained, corresponding to a 80-dimensional feature vector. The CLD represents the spatial layout of the images in a very compact form [7]. It is obtained by applying the discrete cosine transform (DCT) on the 2-D array of local representative colors in $YCbCr$ color space, where Y is the brightness (luma), Cb is blue minus luma ($B - Y$) and Cr is red minus luma ($R - Y$). In this work, CLD with 10 Y , 3 Cb and 3 Cr coefficients is extracted to form a 16-dimensional feature vector.

To extract the semi-global feature vector, a simple grid-based (4×4) approach [6] is used to divide the images into 16 blocks and finally into five overlapping sub-images. Several moment based color and texture features are extracted from each of the sub-images and later they are combined to form a semi-global feature vector. For the moment-based color feature, the first (mean) and second (standard deviation) central moments of each color channel in HSV (Hue, Saturation, and Value) space are extracted. For the texture features, second order moments, such as energy, maximum probability, entropy, contrast and inverse difference are extracted from the grey level co-occurrence matrix (GLCM) [8].

Color and texture feature vectors are normalized and combined to form a joint feature vector of 11-dimensions (6 for color and 5 for texture) for each sub-region to finally generate a 55-dimensional (5×11) semi-global feature vector.

Since images in the different collections vary in sizes, resizing them into a thumbnail of a fixed size might reduce some noise due to the presence of artifacts in the images (specially in medical images), although it might introduce distortion. Hence, each image is converted to gray-level and down scaled to 64×64 regardless of the original aspect ratio. Next, the down-scaled image is partitioned further with a 16×16 grid to form small blocks of (4×4) pixels. The average gray value of each block is measured and concatenated to form a 256-dimensional feature vector. The average gray value of each block is measured to compensate for the global or local image deformations to some extent and add robustness with respect to translation and intensity changes.

We also considered a local region specific feature extraction approach based on our previous work in [6]. In this approach, each image is automatically segmented into a set of homogeneous regions based on a fast K-means clustering technique. To represent each region, average color and covariances of each channel are considered. Color feature of each region is a 3-D vector and is represented by the K-means cluster center, i.e., the average value for each color channel in HSV space of all the image pixels in this region. Texture feature of each region is measured in an indirect way by considering the cross-correlation among color channels due to the off-diagonal of the 3×3 covariance matrix of the region. Finally, the local region specific features of each image are utilized in an image level similarity matching function by considering individual region level distance measures at first [6].

The overall image level similarity is measured by fusion of a weighted combination of individual similarity measures of different image representation as discussed above. For the query image Q and the target image T in the database, the similarity measures of global, semi-global, scaled, and local features are defined as $S_g(Q, T)$, $S_{sg}(Q, T)$, $S_s(Q, T)$, and $S_l(Q, T)$ respectively. For the global, semi-global and scaled specific features, the weighted Euclidean-based similarity measures are utilized as is the Bhattacharyya distance measure [9] for the local feature. The details of the similarity matching functions can be found in [6]. Finally, the similarity matching functions are linearly combined as follows:

$$S_{\text{image}}(Q, T) = w_g S_g(Q, T) + w_{sg} S_{sg}(Q, T) + w_s S_s(Q, T) + w_l S_l(Q, T) \quad (1)$$

Here, w_g , w_{sg} , w_s and w_l are non-negative weighting factors of different feature level similarities, with $w_g + w_{sg} + w_s + w_l = 1$.

2.2 Keyword-Based Retrieval Approach

The indexing technique for keyword search is based on the popular vector space model of information retrieval (IR) [4]. In this model, texts and queries are represented as vectors in a N -dimensional space, where N is the number of keywords in the collection. Each document j is represented as a vector: $\mathbf{D}_j = \langle$

$w_{1j}, \dots, w_{Nj} > 0$. The element w_{ij} , $i \in (1, \dots, N)$ represents the weights of the keyword w_i appearing in document j and both the local and global weights are accounted for by considering the TF-IDF approach [4]. The local weight is denoted as $L_{ij} = \log(f_{ij}) + 1$, where f_{ij} is the frequency of occurrence of keyword w_i in document j . The global weight is denoted by an inverse document frequency as $G_i = \log(M/M_i) + 1$, where M_i is the number of documents in which w_i is found and M is the total number of documents in the collection. The element w_{ij} is expressed as the product of the local and global weights as $w_{ij} = L_{ij} * G_i$. Finally, the cosine similarity measure is adopted between the feature vectors of the query q and database document j as follows [4]:

$$S_{\text{text}}(q, j) = S_{\text{text}}(\mathbf{D}_q, \mathbf{D}_j) = \frac{\sum_{i=1}^N w_{iq} * w_{ij}}{\sqrt{\sum_{i=1}^N (w_{iq})^2} * \sqrt{\sum_{i=1}^N (w_{ij})^2}} \quad (2)$$

where, $\mathbf{D}_q$ and $\mathbf{D}_j$ are the query and document vectors respectively.

2.3 Cross-Modal Interaction with Relevance Feedback

In a practical multi-modal image retrieval system, the user at first might want to search images with keywords as it is more convenient and semantically more appropriate. However, a short query (e.g., query topic) with few keywords might not be enough to incorporate the user perceived semantics to the retrieval system. In this paper, a simpler approach of query expansion is considered based on identifying useful terms or keywords from the associated annotation files for the images. The approach of the textual query expansion based on the RF is as follows: the user provides the initial query topic and the system extracts from it a set of keywords to form the initial textual query vector $\mathbf{D}_{q(0)}$. This query vector is used to retrieve the K most similar images from the associated annotation files. If the user is not satisfied with the result, then the system allows him/her to select a set of relevant (positive) images close to the semantics of the initial textual query topic. In the next iteration, it extracts the most frequently occurring ten keywords from the relevant annotation files. After extracting additional keywords, the query vector is updated as $\mathbf{D}_{q(i)}$ at iteration i by re-weighting its keywords based on the TF-IDF weighting scheme and is re-submitted to the system as the new query. This process continues for several iterations until the user is satisfied with the text only result.

The visual features of the images also play an important role in distinguishing different image categories. Therefore, we also need to perform RF with the content-based image search for precision [3]. In this scenario, as in the textual query expansion, the user might provide the initial image query vector $\mathbf{f}_{Q(0)}$ to retrieve the K most similar images based on (1). In the next iterations, the user selects a set of relevant (positive) images which are quite similar to the initial query image. It is assumed that, all the positive feedback images at some

particular iteration i will belong to the user perceived semantic category and obey the Gaussian distribution to form a cluster in the feature space.

Let, N_{Pos} be the number of positive feedback images at iteration i and $\mathbf{f}_{T_j} \in \mathbb{R}^d$ be the feature vector that represents j -th image for $j \in \{1, \dots, N_{\text{Pos}}\}$, then the new query point at iteration i is estimated as

$$\mathbf{f}_{Q(i)} = \frac{1}{N_{\text{Pos}}} \sum_{j=1}^{N_{\text{Pos}}} \mathbf{f}_{T_j} \quad (3)$$

as the mean vector of the positive images and the covariance matrix is estimated as

$$\mathbf{C}_{(i)} = \frac{1}{N_{\text{Pos}} - 1} \sum_{j=1}^{N_{\text{Pos}}} (\mathbf{f}_{T_j} - \mathbf{f}_{Q(i)})(\mathbf{f}_{T_j} - \mathbf{f}_{Q(i)})^T \quad (4)$$

However, the singularity issue might arise in the covariance matrix estimation if fewer than $d + 1$ (where d is the dimension of a particular feature vector) samples or positive images are available (as it will always be the case for the number of user feedback images). So, we add the following regularization to avoid singularity in matrices as [10]:

$$\hat{\mathbf{C}}_{(i)} = \alpha \mathbf{C}_{(i)} + (1 - \alpha) \mathbf{I} \quad (5)$$

for $0 \leq \alpha \leq 1$ and $\mathbf{I}$ is the $d \times d$ identity matrix. For the experiments, we select the value of α as 0.8. After generating the mean vector and covariance matrix of the positive images, we adjust or change the Euclidean distance measures of various feature representations with the Mahalanobis distance measure [9] in the successive iterations as :

$$\text{DIS}_{\text{Maha}}(Q, T) = (\mathbf{f}_{Q(i)} - \mathbf{f}_T)^T \hat{\mathbf{C}}_{(i)}^{-1} (\mathbf{f}_{Q(i)} - \mathbf{f}_T) \quad (6)$$

Here, $\mathbf{f}_T$ denotes the feature vector of the database image T in general for the different image representations. The Mahalanobis distance differs from the Euclidean distance in that it takes into account the correlations of the feature attributes of the relevant images. Basically, at each iteration of the RF, we estimate the mean vectors and covariance matrices for the global, semi-global and scaled feature vectors separately and utilize them in the Mahalanobis distance measures. We did not consider the region specific local feature in the RF scheme. Finally, a ranked result list is obtained by applying (1) with the modified individual distance measures.

2.4 Integration of Textual and Visual Results

We considered a pre-filtering and merging approach based on the results obtained by applying RF on the textual and visual only searches. The steps involved in the proposed interaction and integration approaches are as follows:

Step 1: Perform an initial text-based search with a query vector $\mathbf{D}_{q(0)}$ for a query topic $q(0)$ at iteration $i = 0$ and rank the associated images based on the ranking of the annotation files by applying (2).

Step 2: Consider the top $K = 30$ most similar images from the retrieval interface and obtain user feedback about relevant (positive) images (e.g., associated annotation files) for the textual query expansion.

Step 3: Resubmit the modified query vector $\mathbf{D}_{q(i)}$ by re-weighting the keywords at iteration i . Continue the iterations by going to Step 2 until the user is satisfied or go to Step 4.

Step 4: Perform the visual only search on the result list of the first $L = 2000$ images obtained from step 3 with the initial query image $Q(0)$.

Step 5: Obtain the user feedback of the relevant (positive) images and perform the image only search with the new query $Q(i)$ at iteration i . Continue the iterations by going to Step 4 until the user is satisfied or go to step 6.

Step 6: Finally, linearly combine the text and image based scores as:

$$S(Q, T) = w_{\text{text}}S_{\text{text}}(Q, T) + w_{\text{image}}S_{\text{image}}(Q, T) \quad (7)$$

and rank the images in descending order of similarity values to return the top 1000 images. Here, w_{image} and w_{text} are non-negative weighting factors for image and text based similarity measures, where $w_{\text{image}} + w_{\text{text}} = 1$.

2.5 Analysis of the Retrieval Results

We submitted three different runs for the ad-hoc retrieval of the IAPR collection and three runs for the ad-hoc retrieval of the medical collection as shown in Table 1 and Table 2. In all these runs, only English is used as the source and target language without any translation. For image only retrieval, the feature level weights are selected as $w_g = 0.4$, $w_{sg} = 0.3$, $w_s = 0.1$ and $w_l = 0.2$, whereas for the final merging, the weights are selected as $w_{\text{text}} = 0.7$ and $w_{\text{image}} = 0.3$ for the text and image-based search.

In the first run with ID “*Cindi-Text-Eng*”, we performed only the automatic text-based search without any feedback as our base run. For the second and third runs with ID “*Cindi-TXT-EXP*” and “*Cindi-Exp-RF*” respectively, we performed manual feedback in the text only modality and both text and image modalities (with only one or two iterations for each modality) as discussed in the previous sections. From Table 1, it is clear that the mean average precision (MAP) scores are almost doubled in both cases and the best performance is obtained with feedback and integration of text and image search. In fact, these two runs ranked first and second in terms of the MAP score among the 157 submissions in the photographic retrieval task. Our group performed manual submissions using relevance judgement from the user. This along with the integration of both modalities could be the reason for the good results. For the medical retrieval task, we performed only the automatic visual only search without feedback as shown in Table 2 for the first run with ID “*CINDI-Fusion-Visual*”. Our group ranked first in this run in the category (automatic+visual) based

Table 1. Results of the ImageCLEFphoto retrieval task

Run ID	Modality	A/M	RF	QE MAP
Cindi-Text-Eng	Text	Automatic	Without	No 0.1995
Cindi-TXT-EXP	Text	Manual	With	Yes 0.3749
Cindi-Exp-RF	Text+Image	Manual	With	Yes 0.3850

on the MAP score (0.0753) out of 11 different submissions. For the second run with ID “*CINDI-Visual-RF*”, we performed manual feedback in the image only modality. For this category (e.g., visual only run with RF), only our group has participated this year and achieved a better MAP score (0.0957) than without RF as shown in Table 2. For the third run with ID “*CINDI-Text-Visual-RF*”, we performed manual feedback in both the modalities and merged the result lists as discussed in the previous section. For this category (e.g., mixed with RF), with the MAP score (0.1513), it is clear that combining both modalities is far better than using a single one.

Table 2. Results of the medical retrieval task

Run ID	Modality	A/M	RF	QE MAP
CINDI-Fusion-Visual	Image	Automatic	Without	No 0.0753
CINDI-Visual-RF	Image	Manual	With	No 0.0957
CINDI-Text-Visual-RF	Image+Text	Manual	With	Yes 0.1513

3 Automatic Annotation Tasks

The aim of the automatic annotation task is to compare the state of the art approach to image classification and annotation, and to quantify the improvement for image retrieval. We investigate a supervised learning approach to associate the low-level image features with their high-level semantic categories for the image categorization or annotation of the medical (IRMA) and object (LTU) data sets. Specially, we explore a classifier combination approach of using several probabilistic multi-class support vector machine (SVM) classifiers. Thus, instead of using only one integrated feature vector, we utilize features at different levels of image representation as inputs to the SVM classifiers and use several classifier combination rules (such as product, sum, max, etc.) to predict the final image category, as well as, probability or membership score of each category as the image annotation.

SVM was originally designed for binary classification problems [11]. It performs classification between two classes by finding a decision surface that is based on the most informative points of the training set. A number of methods have been proposed for extension to multi-class problem to separate L mutually exclusive classes by essentially solving many two-class problems and combining their predictions in various ways [12]. In the experiments, we utilize a multi-class classification method

by combining all pairwise comparisons of binary SVM classifiers, known as *one-against-one* or pairwise coupling (PWC) [12]. PWC constructs binary SVM's between all possible pairs of classes. Hence, this method uses $L*(L-1)/2$ binary classifiers, each of which provides a partial decision for classifying a data point. During the testing of a feature vector $\mathbf{f}$, each of the $L*(L-1)/2$ classifier votes for one class. The winning class is the one with the largest number of accumulated votes. Although the voting procedure requires just pairwise decisions, it only predicts a class label [14]. However, to annotate or represent each image with a category specific confidence score, a probability estimation is required. In our experiments, the probability estimation approach in [12] for the multi-class classification by PWC is utilized. In this context, given the observation or feature vector $\mathbf{f}$, the goal is to estimate the posterior probability as

$$p_k = P(y = k \mid \mathbf{f}), k = 1, \dots, L \quad (8)$$

3.1 Multiple SVM Classifier Combination

In general, classifier combination is defined as the instances of classifiers with different structures trained on distinct feature spaces [13]. Feature descriptors at different levels of image representation are in diversified forms and are often complementary in nature. Hence, multiple classifiers are needed to deal with different features. In the experiments, we consider expert combination strategies of the SVM classifiers with different low-level features as inputs based on three popular classifier combination rules (e.g., sum, product and max rules) [13]. Since the output of the classifiers is to be combined, the confidence or membership scores from the probabilistic SVM's in the range of $[0, 1]$ for each category serve this purpose. The multi-class SVM classifiers as experts on different feature descriptors, as described in Sect. 2.1 for the retrieval task, are combined with the above rules to classify the images to a category with the highest obtained probability value and annotate the images with the probability or membership scores as shown in the process diagram in Fig. 1.

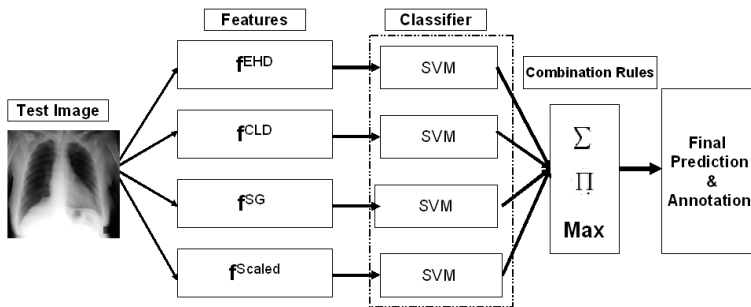


Fig. 1. Block diagram of the classifier combination process

3.2 Analysis of the Annotation Results

To perform SVM-based classification, we utilized the LIBSVM software package [14]. For the training of both the data sets, radial basis functions (RBF) as kernels are utilized with different γ and C parameters; these parameters were determined experimentally with a 5-fold cross validation (CV). We submitted four runs for the object annotation task as shown in Table 3. The first three of these runs are results of experiments with a classifier combination approach with different features as inputs and the last run with ID “*Cindi-Fusion-Knn*” is result of the experiments with a K-NN (K=9) classifier by using the fusion-based similarity matching function in (1). Our best run (e.g., “*Cindi-SVM-Product*”) in this task ranked third among all the submissions, although the accuracy rate is much lower due to the complexity of the data set in general. For the medical

Table 3. Results of the object annotation task

Run ID	Feature	Method	Error rate(%)
Cindi-SVM-Product	CLD+EHD+Semi-global	SVM (Product)	83.2
Cindi-SVM-SUM	CLD+EHD+Semi-global	SVM (Sum)	85.2
Cindi-SVM-EHD	EHD	SVM	85.0
Cindi-Fusion-Knn	CLD+EHD+Semi-global+Local	K-NN	87.1

annotation task, we submitted five runs as shown in Table 4. The first four of these runs are the result of experiments with the multi-class SVM and classifier combination approach and the last run with ID “*cindi-fusion-KNN9*” is the result of experiments with a K-NN (K=9) classifier. Our best run (e.g., “*cindi-svm-sum*”) in this task ranked 13th among all the submissions and 6th among all the groups.

Table 4. Results of the medical annotation task

Run ID	Feature	Method	Error rate(%)
cindi-svm-product	CLD+EHD+Scaled+Semi-global	SVM (Product)	24.8
cindi-svm-sum	CLD+EHD+Scaled+Semi-global	SVM (Sum)	24.1
cindi-svm-max	CLD+EHD+Scaled+Semi-global	SVM (Max)	26.1
cindi-svm-ehd	EHD	SVM	25.5
cindi-fusion-KNN9	CLD+EHD+Scaled+Semi-global+Local	K-NN	25.6

4 Conclusion

This paper presents the image retrieval and annotation approach used by the CINDI research group for ImageCLEF 2006. We participated in all four sub-tasks and submitted several runs with different combination of methods, features and

parameters. We experimented with a cross-modal interaction and integration approach for the retrieval of the photographic and medical image collections and a supervised classifier combination-based approach for the automatic annotation of the object and medical data sets. The analysis and the results of the runs are discussed in this paper.

References

1. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In: Seventh Workshop of the Cross-Language Evaluation Forum. Alicante, Spain. LNCS (to appear, 2006)
2. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersch, W.: Overview of the ImageCLEFmed, medical retrieval and annotation tasks. In: Seventh Workshop of the Cross-Language Evaluation Forum. Alicante, Spain, LNCS (to appear, 2006)
3. Rui, Y., Huang, T.S.: Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Circuits Syst. Video Technol.* 8(5), 644–655 (1999)
4. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley, Reading (1999)
5. Smeulder, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. on Pattern Anal. and Machine Intell.* 22, 1349–1380 (2000)
6. Rahman, M.M., Desai, B.C., Bhattacharya, P.: A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval. In: *Proc. 9th Internat Conf. Information Fusion* (2006)
7. Manjunath, B.S., Salembier, P., Sikora, T. (eds.): *Introduction to MPEG-7 – Multimedia Content Description Interface*, pp. 187–212. John Wiley Sons Ltd. Chichester (2002)
8. Haralick, R.M., Shanmugam, D.I.: Textural features for image classification. *IEEE Trans System, Man, Cybernetics SMC-3*, 610–621 (1973)
9. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, 2nd edn. Academic Press Professional, Inc. San Diego, CA, USA (1990)
10. Friedman, J.: Regularized Discriminant Analysis. *Journal of American Statistical Association* 84, 165–175 (2002)
11. Vapnik, V.: *Statistical Learning Theory*. Wiley, New York (1998)
12. Wu, T.F., Lin, C.J., Weng, R.C.: Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research*. 5, 975–1005 (2004)
13. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. *IEEE Trans Pattern Anal Machine Intell.* 20 (3), 226–239 (1998)
14. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines, Software (2001), available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

Image Retrieval and Annotation Using Maximum Entropy

Thomas Deselaers, Tobias Weyand, and Hermann Ney

Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
`deselaers@cs.rwth-aachen.de`

Abstract. We present and discuss our participation in the four tasks of the ImageCLEF 2006 Evaluation. In particular, we present a novel approach to learn feature weights in our content-based image retrieval system FIRE. Given a set of training images with known relevance among each other, the retrieval task is reformulated as a classification task and then the weights to combine a set of features are trained discriminatively using the maximum entropy framework. Experimental results for the medical retrieval task show large improvements over heuristically chosen weights. Furthermore the maximum entropy approach is used for the automatic image annotation tasks in combination with a part-based object model. Using our object classification methods, we obtained the best results in the medical and in the object annotation task.

1 Introduction

An important task for obtaining satisfying results in content-based image retrieval and annotation is the combination and weighting of descriptors extracted from the images. Commonly, machine learning algorithms are used to learn a good way of combining descriptors from some training data. Discriminative log-linear models [1] have been successfully used for this purpose in different scenarios: natural language processing [2,1], data mining [3], and image recognition [4,5].

Here, we present, how we used the maximum entropy approach to train feature weights in our content-based image retrieval system for the medical image retrieval task in ImageCLEF 2006. We furthermore present our results of the ImageCLEF 2006 photographic retrieval task, and of the ImageCLEF 2006 medical annotation and object recognition tasks. In the annotation and recognition tasks, the maximum entropy method was also extensively used and excellent results were obtained.

2 Retrieval Tasks

ImageCLEF 2006 hosted two independent retrieval tasks: The medical retrieval task [6] and the photo retrieval task [7].

For the retrieval tasks the FIRE image retrieval system¹ was used. FIRE is a research image retrieval system that was designed with extensibility in mind and allows to combine various image descriptors and comparison measures easily. It also allows for combining textual information and meta data information with content-based queries.

2.1 FIRE – The Flexible Image Retrieval System

Given a set of positive example images Q^+ and a (possibly empty) set of negative example images Q^- a score $S(Q^+, Q^-, X)$ is calculated for each image X from the database:

$$S(Q^+, Q^-, X) = \sum_{q \in Q^+} S(q, X) + \sum_{q \in Q^-} (1 - S(q, X)). \quad (1)$$

where $S(q, X)$ is the score of database image X with respect to query q and is calculated as $S(q, X) = e^{-\gamma D(q, X)}$ with $\gamma = 1.0$. $D(q, X)$ is a weighted sum of distances calculated as

$$D(q, X) := \sum_{m=1}^M w_m \cdot d_m(q_m, X_m). \quad (2)$$

Here, q_m and X_m are the m th feature of the query image q and the database image X , respectively. d_m is the corresponding distance measure and w_m is a weighting coefficient. For each d_m , $\sum_{X \in \mathcal{B}} d_m(Q_m, X_m) = 1$ is enforced by re-normalization.

Given a query (Q^+, Q^-) , the images are ranked according to descending score and the K images X with highest scores $S(Q^+, Q^-, X)$ are returned by the retriever.

The selection of the weights w_m is a critical step which has so far been done heuristically based on experiences from earlier experiments. However, in the next section we describe, how the maximum entropy framework can be used to learn weights from training data.

2.2 Maximum Entropy Training for Image Retrieval

To obtain suitable feature weights, the maximum entropy approach is promising, because it is ideally suited to combine features of different types and it leads to good results in other areas as mentioned above.

We consider the problem of image retrieval to be a classification problem. Given the query image, the images from the database have to be classified to be either relevant (denoted by $\oplus$) or irrelevant (denoted by $\ominus$). As classification method we choose log-linear models that are trained using the maximum entropy criterion and the GIS algorithm [8].

¹ <http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html>

As features f_i for the log-linear models we choose the distances between the m -th feature of the query image Q and the database image X :

$$f_i(Q, X) := d_i(Q_i, X_i).$$

To allow for prior probabilities, we include a constant feature $f_{i=0}(Q, X) = 1$. Then, the scores $S(q, X)$ from Eq. (1) are replaced by the posterior probability for class $\oplus$ and the ranking and combination of several query images is done as before:

$$\begin{aligned} S(q, X) &:= p(\oplus|Q, X) \\ &= \frac{\exp[\sum_i \lambda_{\oplus i} f_i(Q, X)]}{\sum_{k \in \{\oplus, \ominus\}} \exp[\sum_i \lambda_{ki} f_i(Q, X)]} \end{aligned} \quad (3)$$

Alternatively, Eq. 3 can easily be transformed to be of the form of Eq. 1 and the w_i can be expressed as a function of $\lambda_{\oplus i}$ and $\lambda_{\ominus i}$.

In addition to considering only the first order features as they are described above, we propose to use supplementary second order features (i.e. products of distances) as this usually yields superior performance on other tasks. Given a query image Q and a database image X we use the following set of features:

$$\begin{aligned} f_i(Q, X) &:= d_i(Q_i, X_i) \\ f_{i,j}(Q, X) &:= d_i(Q_i, X_i) \cdot d_j(Q_j, X_j), \quad i \geq j, \end{aligned}$$

again including the constant feature $f_{i=0}(Q, X) = 1$ to allow for prior probabilities. The increased number of features results in more parameters to be trained. In experiments in other domains, features of higher degree have been tested and not found to improve the results, and thus we did not try higher order features.

In the training process, the values of the λ_{ki} are optimized. A sufficiently large amount of training data is necessary to do so. We are given the database $\mathcal{T} = \{X_1, \dots, X_N\}$ of training images with known relevances. For each image X_n we are given a set $R_n = \{Y \mid Y \in \mathcal{T} \text{ is relevant, if } X_n \text{ is the query.}\}$.

Because we want to classify the relation between images into the two categories “relevant” or “irrelevant” on the basis of the distances between their features, we choose the following way to derive the training data for the GIS algorithm: The distance vectors $D(X_n, X_m) = (d_1(X_{n1}, X_{m1}), \dots, d_I(X_{nI}, X_{mI}))$ are calculated for each pair of images $(X_n, X_m) \in \mathcal{T} \times \mathcal{T}$. That is, we obtain N distance vectors for each of the images X_n . These distance vectors are then labeled according to the relevances: Those $D(X_n, X_m)$ where X_m is relevant with respect to X_n , i.e. $X_m \in R_n$, are labeled $\oplus$ (relevant) and the remaining ones are labeled with the class label $\ominus$ (irrelevant).

Given these N^2 distance vectors and their classification into “relevant” and “irrelevant” we train the λ_{ki} of the log-linear model from Eq. (3) using the GIS algorithm.

Table 1. Summary of our runs submitted to the medical retrieval task. The numbers give the weights (empty means 0) of the features in the experiments and the columns denote: *En*: English text, *Fr*: French text, *Ge*: German text, *CH*: color histogram, *GH*: gray histogram, *GTF*: global texture feature, *IH*: invariant feature histogram, *TH*: Tamura Texture Feature histogram, *TN*: 32x32 thumbnail, *PH*: patch histogram. The first group of experiments uses only textual information, the second group uses only visual information, the third group uses textual and visual information, and the last group both types of information and the weights are trained using the maximum entropy approach. The last column gives the results of the evaluation. The last three lines are unsubmitted runs that were performed after the evaluation ended.

run-tag	En	Fr	Ge	CH	GH	GTF	IFH	TH	TN	PH	MAP
En	1										0.15
SimpleUni				1	1	1	1	1	1		0.05
Patch										1	0.04
IfhTamThu							2	2	1		0.05
EnIfhTamThu	1						2	2	1		0.09
EnFrGeIfhTamThu	2	1	1				2	2	1		0.13
EnFrGePatches	2	1	1							1	0.17
EnFrGePatches2	2	1	1							2	0.16
ME [500 iterations]	*	*	*	*	*	*	*	*	*	0	0.07
ME [5000 iterations]	*	*	*	*	*	*	*	*	*	0	0.15
ME [10000 iterations]	*	*	*	*	*	*	*	*	*	0	0.18
ME [20000 iterations]	*	*	*	*	*	*	*	*	*	0	0.18

2.3 Medical Retrieval Task

We submitted nine runs to the medical retrieval task [6], one of these using only text, three using only visual information, and five using visual and textual information. For one of the combined runs we used the above-described maximum entropy training method. To determine the weights, we used the queries and their qrels from last year's medical retrieval task as training data. Table 1 gives an overview of the runs we submitted to the medical retrieval task and the results obtained.

In Figure 1 the trained feature weights are visualized after different numbers of maximum entropy training iterations. It can clearly be seen that after 500 iterations the weights hardly differ from uniform weighting and that thus not enough training iterations were performed. After 5000 iterations, there is a clear gain in performance (cf. Table 1) and the weights are not uniform any more. For example, the weight for feature 1 (English text) has the highest weight. With more iterations, the differences between the particular weights become bigger; after 10.000 iterations no additional gain in performance is yielded anymore.

2.4 Photo/Ad-Hoc Retrieval Task

For the photo- and the ad-hoc retrieval task the newly created IAPR TC-12 database [9] was used, which currently consists of 20,000 general photographs,

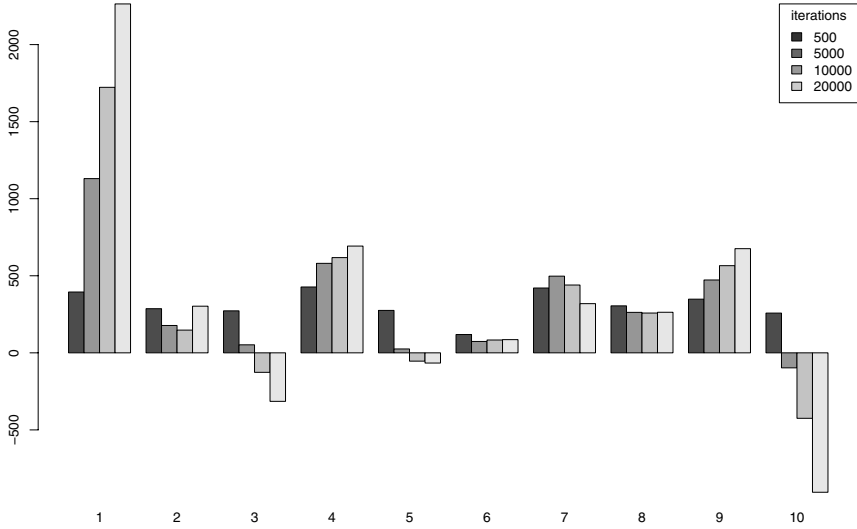


Fig. 1. Trained weights for the medical retrieval task after different numbers of iterations in the maximum entropy training. On the x-axis, the features are given in the same order as in Table 1 and on the y-axis $\lambda_{\oplus i} - \lambda_{\ominus i}$ is given.

mainly from a vacation domain. For each of the images a German and an English description exists. The tasks are described in detail in [7].

Two tasks were defined on this dataset: An ad-hoc task of 60 queries of different semantic and syntactic difficulty, and a photo task of 30 queries, which was based on a subset aiming to investigate the possibilities of purely visual retrieval. Therefore, some semantic constraints were removed from the queries. All queries were formulated by a short textual description and three positive example images.

Due to short time, we were unable to tune any parameters and just chose to submit two purely visual, full-automatic runs to these tasks.

In Table 2 we summarize the outcomes of the two tasks using the IAPR TC-12 database. The overall MAP values are rather low, but the combination of invariant feature histograms and Tamura texture features clearly outperforms all competing methods.

For the runs entitled **IFHTAM**, we used a combination of invariant feature histograms and Tamura texture histograms. Both histograms are combined by Jeffrey divergence and the invariant feature histograms are weighted by a factor of 2. This combination has been seen to be a very effective combination of features for databases of general photographs like for example the Corel database [10]. For the runs entitled **PatchHisto** we used histograms of vector-quantized image patches with 2048 bins.

All runs we submitted were top-ranked in the category “visual retrieval, no user interaction”. A detailed analysis of the results is given in [7].

Table 2. Results from the AdHoc and the Photo task

(a) Results from the adhoc retrieval task with 60 queries in the category “visual only, full automatic, no user interaction”.				(b) Results from the photo retrieval task with 30 queries. All submissions to this task were submitted as full automatic, visual only submissions without user feedback.			
task	run-tag	map	rank	task	run-tag	map	rank
RWTHi6	IFHTAM	0.06	1	RWTHi6	IFHTAM	0.11	1
RWTHi6	PatchHisto	0.05	2	RWTHi6	PatchHisto	0.08	2
CEA	mPHic	0.05	3	IPAL	LSA3	0.07	3
CEA	2mPHit	0.04	4	IPAL	LSA2	0.06	5
IPAL	LSA	0.03	5	IPAL	LSA1	0.06	4
IPAL	MF	0.02	6	IPAL	MF	0.04	6

3 Automatic Annotation Tasks

The following sections describe the methods we applied to the automatic annotation tasks and the experiments we performed.

3.1 Image Distortion Model

The image distortion model [11,12] is a zeroth-order image deformation model to compare images pixel-wise. Here, classification is done using the nearest neighbor decision rule: to classify an image, it is compared to all training images in the database and the class of the most similar image is chosen. To compare images, the Euclidean distance can be seen as a very basic baseline, and in earlier works it was shown that image deformation models are a suitable way to improve classification performance significantly e.g. for medical radiographs and for optical character recognition [12,11]. Here we allow each pixel of the database images to be aligned to the pixels from a 5×5 neighborhood from the image to be classified taking into account the local context from a 3×3 Sobel neighborhood.

This method is of particular interest as it outperformed all other methods in automatic annotation task of ImageCLEF 2005 [13].

3.2 Sparse Patch Histograms and Discriminative Classification

This approach is based on the widely adopted assumption that objects in images can be represented as a set of loosely coupled parts. In contrast to former models [14], this method can cope with an arbitrary number of object parts. Here, the object parts are modelled by image patches that are extracted at each position and then efficiently stored in a histogram. In addition to the patch appearance, the positions of the extracted patches are considered and provide a significant increase in the recognition performance.

Using this method, we create sparse histograms of 65536 ($2^{16} = 8^4$) bins, which can either be classified using the nearest neighbor rule and a suitable histogram comparison measure or a discriminative model can be trained for classification. Here, we used a support vector machine with a histogram intersection kernel and a discriminatively trained log-linear maximum entropy model.

A detailed description of the method is given in [15].

3.3 Patch Histograms and Maximum Entropy Classification

In object recognition and detection currently the assumption that objects consist of parts that can be modelled independently is very common, which led to a wide variety of bag-of-features approaches [16,14].

Here we follow this approach to generate histograms of image patches for image retrieval. The creation is a 3-step procedure:

1. in the first phase, sub-images are extracted from all training images and the dimensionality is reduced to 40 dimensions using PCA transformation.
2. in the second phase, the sub-images of all training images are jointly clustered using the EM algorithm for Gaussian mixtures to form 2000-8000 clusters.
3. in the third phase, all information about each sub-image is discarded except its closest cluster center. Then, for each image a histogram over the cluster identifiers of the respective patches is created, thus effectively coding which “visual words” from the code-book occur in the image.

These histograms are then classified using the maximum entropy approach [14].

3.4 Medical Automatic Annotation Task

We submitted three runs to the medical automatic annotation task [6]: one run using the image distortion model RWTHi6-IDM, with exactly the same settings as the according run from last year, which clearly outperformed all competing methods [17] and two other runs based on sparse histograms of image patches [15], where we used a discriminatively trained log-linear maximum entropy model (RWTHi6-SHME) and support vector machines with a histogram intersection kernel (RWTHi6-SHSVM) respectively. Due to time constraints we were unable to submit the method described in Section 3.3, but we give comparison results here.

Results. The results of the evaluation are given in detail in the overview paper [6]. Table 3 gives an overview on our submissions and the best competing runs and it can be seen that the runs using the discriminative classifier for the histograms clearly outperform the image distortion model and that in summary our method performed very good on the task.

Concluding it can be seen that the approach, where local image descriptors were extracted at every position in the image, outperformed our other approaches. Probably the modelling of absolute positions is suitable for radiograph

Table 3. An overview of the results of the medical automatic annotation task. The first part gives our results (including the error rate of an unsubmitted method for comparison to the results of last year); the second part gives results from other groups that are interesting for comparison.

rank	run-tag	error rate[%]
1	RWTHi6 SHME	16.2
2	RWTHi6 SHSVM	16.7
11	RWTHi6 IDM	20.5
-	RWTHi6 - [14]	22.4
2	UFR ns1000-20x20x10	16.7
4	MedIC-CISMef local+global-PCA335	17.2
12	RWTHmi rwthmi	21.5
23	ULG sysmod-random-subwindows-ex	29.0

Table 4. Results from the object annotation task

rank	Group	ID run-tag	Error rate
1	RWTHi6	SHME	77.3
2	RWTHi6	PatchHisto	80.2
3	CINDI	SVM-Product	83.2
4	CINDI	SVM-EHD	85.0
5	CINDI	SVM-SUM	85.2
6	CINDI	Fusion-knn	87.1
7	DEU-CS	edgehistogr-centroid	88.2
8	DEU-CS	colorlayout-centroid	93.2

recognition, because it seems to be a suitable assumption that radiographs are taken under controlled conditions and that thus the geometric layout of images showing the same body region can be assumed to be very similar.

3.5 Object Annotation Task

We submitted two runs to this task [7], one using the method with vector quantized histograms described in Section 3.3 (run-tag **PatchHisto**) and the other using the method with sparse histograms as described in Section 3.2 (run-tag **SHME**). These two methods were also used in the PASCAL visual object classes challenge 2006. The third method [18] we submitted to the PASCAL challenge could not be applied to this task due to time and memory constraints.

Results. Table 4 gives the results of the object annotation task. On the average, the error rates are very high. The best two results of 77.3% and 80.2% were achieved with our discriminative classification method. For the submissions of the CINDI group, support vector machines were used and the DEU-CS group used a nearest neighbor classification. Obviously, the results are not satisfactory and large improvements should be possible.

4 Conclusion and Outlook

We presented our efforts for the ImageCLEF 2006 image retrieval and annotation challenge. In particular, we presented a discriminative method to train weights to combine features in our image retrieval system. This method allows to find weights that clearly outperform a setting with feature weights chosen from experiences from earlier experiments and thus allowed us to obtain better results than our best old system. The trained features weights can be interpreted to see which features are most important for a given task and this effect is smoothly achieved by an iterative training procedure.

The maximum entropy principle was furthermore used for automatic image annotation and very well results were obtained.

Acknowledgement

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

References

1. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J.: A maximum entropy approach to natural language processing. *Computational Linguistics* 22, 39–72 (1996)
2. Bender, O., Och, F., Ney, H.: Maximum entropy models for named entity recognition. In: 7th Conference on Computational Natural Language Learning, Edmonton, Canada, pp. 148–152 (2003)
3. Mauser, A., Bezrukov, I., Deselaers, T., Keysers, D.: Predicting customer behavior using naive bayes and maximum entropy – winning the data-mining-cup 2004. In: *Informatiktage 2005 der Gesellschaft für Informatik*, St. Augustin, Germany (Inpress, 2005)
4. Keysers, D., Och, F.J., Ney, H.: Maximum entropy and Gaussian models for image object recognition. In: *Pattern Recognition, 24th DAGM Symposium*, Zürich, Switzerland, pp. 498–506 (2002)
5. Lazebnik, S., Schmid, C., Ponce, J.: A maximum entropy framework for part-based texture and object recognition. In: *IEEE International Conference on Computer Vision (ICCV 05)*, Beijing, China, vol. 1, pp. 832–838 (2005)
6. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersch, W.: Overview of the imageclefmed, medical retrieval and annotation tasks. In: *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*. LNCS, Alicante, Spain (to appear, 2007)
7. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the imageclef, photographic retrieval and object annotation tasks. In: *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*. LNCS, Alicante, Spain (to appear, 2007)
8. Keysers, D., Och, F.J., Ney, H.: Efficient maximum entropy training for statistical object recognition. In: *Informatiktage 2002 der Gesellschaft für Informatik*, Bad Schussenried, Germany pp. 342–345 (2002)

9. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr benchmark: A new evaluation resource for visual information systems. In: LREC 06 OntoImage 2006: Language Resources for Content-Based Image Retrieval, Genoa, Italy (in press, 2006)
10. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval – a quantitative comparison. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) *Pattern Recognition*. LNCS, vol. 3175, pp. 228–236. Springer, Heidelberg (2004)
11. Keysers, D., Gollan, C., Ney, H.: Local context in non-linear deformation models for handwritten character recognition. In: *International Conference on Pattern Recognition*, Cambridge, UK, vol. 4, pp. 511–514 (2004)
12. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non-linear distortion models. In: *Proc. BVM 2004, Bildverarbeitung für die Medizin*, Berlin, Germany, pp. 366–370 (2004)
13. Clough, P., Mueller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., Hersch, W.: The clef 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
14. Deselaers, T., Keysers, D., Ney, H.: Discriminative training for object recognition using image patches. In: *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, vol. 2, pp. 157–162 (2005)
15. Deselaers, T., Hegerath, A., Keysers, D., Ney, H.: Sparse patch-histograms for object classification in cluttered images. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.) *Pattern Recognition*. LNCS, vol. 4174, pp. 202–211. Springer, Heidelberg (2006)
16. Dorkó, G., Schmid, C.: Object class recognition using discriminative local features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (submitted 2004)
17. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in Image-CLEF 2005: Combining content-based image retrieval with textual information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, Springer, Heidelberg (2006)
18. Hegerath, A., Deselaers, T., Ney, H.: Patch-based object recognition using discriminatively trained gaussian mixtures. In: *17th British Machine Vision Conference (BMVC06)*, Edinburgh, UK, vol. 2, pp. 519–528 (2006)

Inter-media Pseudo-relevance Feedback Application to ImageCLEF 2006 Photo Retrieval

Nicolas Maillot, Jean-Pierre Chevallet, and Joo Hwee Lim

IPAL French-Singaporean Joint Lab
Institute for Infocomm Research (I2R)
Centre National de la Recherche Scientifique (CNRS)
21 Heng Mui Keng Terrace - Singapore 119613
{nmaillot, viscjp, joohwee}@i2r.a-star.edu.sg

Abstract. This paper shows how to use a text retrieval and an image retrieval engine in a *cooperative* way. The proposed Inter-Media Pseudo-Relevance Feedback approach shows how the image modality can be used for expanding the text query and significantly improve the performance (MAP) of a multimedia information retrieval system. The document collection used for experiments is provided by the International Association for Pattern Recognition (IAPR) [1]. The proposed approach is completely automatic and has been used to participate to the ImageCLEF 2006 Photo Retrieval task.

1 Introduction

One of the most interesting issues in multimedia information retrieval is to use different modalities (e.g. text, image) in a cooperative way. Our goal is to study *inter-media pseudo-relevance feedback* (between text and image) and so to explore how the output of an image retrieval system can be used for expanding textual queries. The hypothesis is that two images which are visually similar share some common semantics.

Figure 1 illustrates this principle for text query expansion driven by the image modality. In this case, retrieval is achieved in 3 main steps. (1) The initial query is used as an input of the image retrieval engine. (2) The textual annotations associated with the top k documents retrieved by the image retrieval engine are then used for query expansion purposes. (3) After this expansion, the resulting text query is processed by the text retrieval engine to produce the final set of ranked documents. In our experiments, we have set $k = 3$.

This document is structured as follows. Section 2 gives a description of the system used to produce results submitted by IPAL at ImageCLEF 2006 Photo task. Section 3 provides an analysis of the results obtained. We conclude and present future works in section 4.

2 System Description

2.1 Text Indexing and Retrieval

Text indexing and retrieval were mainly achieved by the XIOTA system [2]. Before indexing, linguistic processing is performed at four levels: morpho-syntactic, noun-phrase, named entity and conceptual.

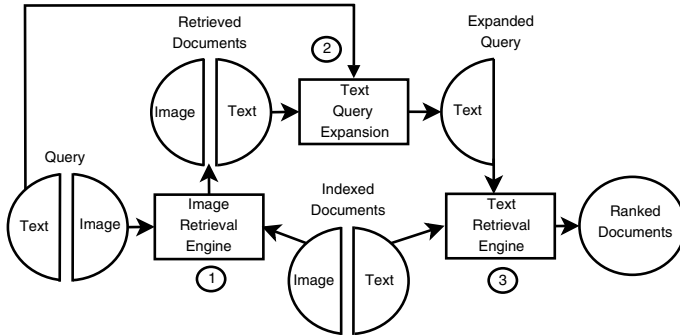


Fig. 1. Overview of *pseudo-relevance* feedback. The textual annotations associated with the top images retrieved are used for expanding the text query.

At the **morpho-syntactic** level, text is processed as follows: Automatic XML correction in order to obtain valid XML documents, part of speech (POS) tagging with TreeTagger¹ [3], Word normalization (e.g. removing of accents), and spelling corrections with aspell². During retrieval, queries are processed in the same way and a classical vector space model is used for matching queries with documents.

Noun Phrase Detection is achieved by using WordNet. Candidates were selected using the following POS templates: *noun (singular) + noun (singular)* (e.g. "baseball cap"); *noun (singular) + noun (plural)* (e.g. "tennis players"); *proper noun + proper noun* (e.g. "South America"); *verb VG + noun* (e.g. "swimming pool"). If the template is matched and present in WordNet, the couple of terms is replaced by one stemmed term.

Named Entity Detection. In this tourist image set, the location of the scene in the picture is important. That is why we have decided to detect geographic Named Entity. We have used two information sources: WordNet and the corpus itself (based on the LOCATION tag).

Concept Indexing tends to reduce the term mismatch because all term variations are replaced by one unique concept. Unfortunately, an issue remains in the concept selection: disambiguation. For this experiment, we have used the WordNet sense frequency by selecting the most frequent meaning (identified by a WordNet concept reference) associated with a term.

2.2 Image Indexing and Retrieval

Feature Extraction. The feature extraction process is based on a tessellation of the images of the database. Each image is split into patches. The visual indexing process is based on patch extraction on all the images of the collection followed by feature extraction on the resulting patches. Let I be the set of the N images in the document collection. First, each image $i \in I$ is split into n_p patches p_i^k ($1 \leq k \leq n_p$). Patch extraction on the whole collection results in $n_p \times N$ patches.

¹ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

² <http://aspell.sourceforge.net/>

The low-level *visual* features extracted on patches are the following: Gabor features [4] and RGBL histograms (32 bins for each component). The resulting feature vectors are respectively dimension 60 and 128.

For a patch p_i^k , the numerical feature vector extracted from p_i^k is noted $fe(p_i^k) \in \mathcal{R}^n$. In this case, $n = 60 + 128$.

We also define a similarity measure based on *regions* obtained by meanshift segmentation [5] as defined in eq. (1):

$$dR(i, j) = \sum_{r_i^k \in R_i} \min\{L_2(fe(r_i^k), fe(r_j^l))\}_{r_j^l \in R_j} / Card(R_i) \quad (1)$$

with $R_i = \{r_i^k\}$, resp. $R_j = \{r_j^l\}$ is the set of regions obtained from segmentation of image i , resp. j .

For a region r_i^k , the numerical feature vector extracted from is noted $fe(r_i^k) \in \mathcal{R}^n$. Additional low-level features extracted on regions are their *size* and the *position* of their centroids.

We have also used Local Features to characterize *fine* details. Note that in this case, patches are not considered. We use bags of SIFT³ as explained in [6]. A visual vocabulary is built by clustering techniques (k-means with $k = 150$). Then, a bag of visterms can be associated with each image of the database. The cosine distance is used to compute the distance between two bags of visterms. The bag of visterms associated with the image i is noted $\mathbf{b}_i \in \mathcal{R}^{150}$.

Similarity Function. The visual similarity between two images i and j , $\delta_I(i, j)$, is defined by eq. (2) (In our experiments $\alpha = 0.4$, $\beta = 0.4$, $\gamma = 0.2$):

$$\delta_I(i, j) = \alpha \times \frac{\sum_{k=1}^{n_p} L_2(fe(p_i^k), fe(p_j^k))}{n_p} + \beta \times \frac{\mathbf{b}_i \cdot \mathbf{b}_j}{\|\mathbf{b}_i\| \|\mathbf{b}_j\|} + \gamma dR(i, j) \quad (2)$$

3 Experimental Results

Mean Average Prevision resulting from each run is summarized in table 1.

Run 1: This run results from the use of the pseudo-relevance feedback described in fig. 1. The textual annotations associated with three images are used for expansion of the text query.

Run 2: This run results from a late fusion (LF) (by a weighted-sum) of the output of the text retrieval engine and the output of the image retrieval engine. MF stands for mixed features (described in section 2.2).

Run 3: P stands for Part of Speech and W for Single word. Only nouns, proper nouns, abbreviations, adjectives and verbs are kept. Stemming is provided by the POS tagger (see section 2.1).

Run 4: Concepts are extracted from WordNet and used to expand documents and queries. Both concepts and original terms are kept in the vector because of the low reliability of disambiguation.

Run 5: This run was produced by the use of features described in section 2.2. The similarity distance used between two images i and j is $\delta_I(i, j)$.

IPAL-PW-PFB3 has produced our best Mean Average Precision. This shows that the inter-media pseudo-relevance feedback approach leads to much better

³ Scale Invariant Feature Transform.

Table 1. Mean Average Precision (MAP) of submitted runs. The best run is produced by pseudo-relevance mechanisms involving the 3 top retrieved images. The best MAP (0.385) in mixed modality was obtained by Concordia University with an approach that requires manual feedback.

Run ID	Modality	MAP	Rank (by modality)
Run 1. IPAL-PW-PFB3	Text and Visual	0.3337	2/43
Run 2. IPAL-WN-MF-LF	Text and Visual	0.0568	42/43
Run 3. IPAL-PW	Text	0.1619	29/108
Run 4. IPAL-WN	Text	0.1428	50/108
Run 5. IPAL-MF	Visual	0.0173	6/6

results than each modality considered separately or even in a late fusion scheme. For textual run only, the use of WordNet concepts decreases the MAP. Note that We have not used any of the semantic links provided by WordNet (e.g. specialization links) and we have noticed some problems in disambiguation (like "church" not recognized as a building but as a community of people).

4 Conclusion and Future Work

Our approach is based on a cooperative approach between an image retrieval and a text retrieval system. These experiments show that the combined use of a text retrieval and an image retrieval systems leads to better performance but only for inter media pseudo relevance feedback and not for late fusion.

The IAPR image database is challenging. Many concepts are represented with a large variety of appearances. Query by content using a few images cannot lead to satisfactory results by using only appearance-based techniques. Indeed, a few samples a given concept cannot capture its conceptual essence. Machine learning techniques should be used to extract concepts from the query images.

References

1. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR benchmark: A new evaluation resource for visual information systems. In: LREC 06 OntoImage 2006: Language Resources for Content-Based Image Retrieval, Genoa, Italy, pp. 13–23 (2006)
2. Chevallet, J.P.: X-IOTA: An open XML framework for IR experimentation. In: Myaeng, S.-H., Zhou, M., Wong, K.-F., Zhang, H.-J. (eds.) AIRS 2004. LNCS, vol. 3411, pp. 263–280. Springer, Heidelberg (2005)
3. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proc. of Int. Conf. on New Methods in Language Processing, pp. 44–49 (1994)
4. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. Pattern Analysis and Machine Intelligence 18, 837–842 (1996)
5. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence 24, 603–619 (2002)
6. Csurka, G., Dance, C., Bray, C., Fan, L., Willamowski, J.: Visual categorization with bags of keypoints. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 1–22. Springer, Heidelberg (2004)

ImageCLEF 2006 Experiments at the Chemnitz Technical University

Thomas Wilhelm and Maximilian Eibl

Technische Universität Chemnitz, Fakultät für Informatik, 09107 Chemnitz, Germany
thomas.wilhelm@s2000.tu-chemnitz.de,
eibl@informatik.tu-chemnitz.de

Abstract. We present a report about our participation in the ImageCLEF photo task 2006 and a short description of our new framework for future use in further CLEF participations. We described and analysed our participation in the monolingual English task. A special Lucene-aligned query expansion and histogram comparison are helping to improve the baseline results.

1 Introduction

Our goal was to develop a new information retrieval framework which can handle different custom retrieval mechanisms (i.e. Lucene, the GNU Image-Finding Tool and others). The framework should be so general that one can use either one of the custom retrieval mechanisms or combine different ones. Therefore our efforts were not focused on better results but to get the framework up and running.

2 System Description

For our participation at CLEF 2006 we constructed a GUI (graphical user interface) quite easy to handle. It bases on Lucene version 1.4 and includes several user-configurable components. For example we wrote an *Analyzer* in which *TokenFilters* can be added and rearranged.

For the participation in ImageCLEF 2006 we needed to extend this system with image retrieval capabilities. But instead of extending the old system we rewrote it and developed a general IR framework. This framework does not necessarily require Lucene and makes it possible to build an abstract layer for other search engines like GIFT (GNU Image-Finding Tool) and others.

The main component of our framework is the class *Run*, which includes all information required to perform and repeat searches. A *Run* contains

- the *Topics*, which can be preprocessed by *TopicFilters*
- an *Index*, which was created by an *Indexer* from a *DataCollection*
- a search method called *Searcher*, which can read and search the *Index*
- and *HitSets*, which contain the results.

To use another search engine one will have to extend the abstract classes *Indexer* and *Searcher*. To use another data collection type (e.g. GIRT4, IAPR, ...) one will have to extend the abstract class *DataCollection*.

In order to merge results there is a special extension of the class *Searcher*: the *MergeSearcher*. It combines several *Searchers* and uses an abstract class *Merger* to merge their results.

2.1 Lucene

The text search engine used is Lucene version 1.9 with an adapted *Analyzer*. As described above we create a *LuceneIndexer* and an abstract *LuceneSearcher* that provides basic abilities to underlying *Searchers* like a selectable Lucene *Analyzer*. The *Analyzer* is based on the *StandardTokenizer* and *LowerCaseFilter* by Lucene and is enhanced by a custom Snowball filter for stemming. We also added a positional stop word filter and used the stop word list suggested on the CLEF website. The *Analyzer* details (stemmer and stop word list) can be configured through a GUI. The *LuceneStandardSearcher* works with the Lucene *MultiFieldQueryParser* and searches in specified fields of the index.

2.2 GIFT (GNU Image-Finding Tool)

The GIFT was not used for a submitted run but it should be mentioned here to show the abilities of the framework. We only implemented a *Searcher*, which can access a GIFT server and takes images to generate queries by example. The underlying index must be created with the command line utility provided by the GIFT.

2.3 IAPR Data Collection

To access the IAPR data collection of ImageCLEF we implemented the class *IaprDataCollection* that provides two special options for the *DataDocuments* to contain special data:

- option 1: the *DataDocuments* contain the annotations
- option 2: the *DataDocuments* contain the histograms

In both cases the associated images can be retrieved with the help of the *DataDocument*.

3 Description of the Runs Submitted

We focused on the monolingual task using the English topics and annotations.

The Lucene index of the annotations contains all fields. All fields are indexed and tokenized except DATE, which is not tokenized. The DOCNO is the only field which is stored, resulting in a quite small index (only 2 MB).

We used the default *BooleanQuery* from Lucene with no further weighting of any field pairs in order to avoid any corpus specific weighting.

The default Lucene search classes (default similarity, etc.) were used to search the topic-annotation field pairs as shown in table 1.

Table 1. Topic-annotation field pairs

Topic field	Annotations searched
TITLE	TITLE, LOCATION, DESCRIPTION, NOTES
NARR	DESCRIPTION, NOTES

Our mechanism of query expansion is adapted to the Lucene-index and is, though simple, a little bit slow. First of all a list of assumable relevant documents is identified. Then the Lucene-generated index is checked for terms also contained in documents in this list. These terms, their frequency and the number of documents in which they are found are stored in a second list. Frequency is used to apply a threshold: only terms with a high frequency are used to expand the query. The new terms are weighted by the ratio of its frequency and the number of containing documents.

For all runs, except “tucEEANT”, the query expansion was applied once, using the first 20 results of the original query.

In the run “tucEEAFT2” the query expansion is applied twice. The first time using the annotations of the example images and the second time using the first 20 results of the original query.

To speed up the colour histogram comparison, we also created an index for the histogram data. For each image the histograms for the three components hue, saturation and brightness are stored and can be retrieved by the DOCNO. Though we use a Lucene index, it is used more like a database.

For the run “tucEEAFTI” all results of the textual search are scored using the Euclidean distance of their colour histograms. The histogram distance makes 30% of the new score. The remaining 70% are taken from the original score.

4 Results

Table 2 shows our results at ImageCLEF 2006. To compare our results, the top four runs of the monolingual English task are included in table 2.

Table 2. Ranks of our submitted runs and the first four positions

Rank EN-EN	Rank ALL	System	MOD	A/M	FB	QE	MAP
1	1	Cindi_Exp_RF	MIXED	MANUAL	WITH	YES	0,3850
2	2	Cindi_TXT_EXP	TEXT	MANUAL	WITH	YES	0,3749
3	3	IPAL-PW-PFB3	MIXED	AUTO	WITHOUT		0,3337
4	5	NTU-EN-EN-AUTO-FB-TXTIMG	MIXED	AUTO	WITH		0,2950

Table 2. *(Continued)*

Rank EN-EN	Rank ALL	System	MOD	A/M	FB	QE	MAP
5	16	tucEEAFT2	TEXT	AUTO	WITH	YES	0,2436
17	33	tucEEAFTI	MIXED	AUTO	WITH	YES	0,1856
18	34	tucEEAFT	TEXT	AUTO	WITH	YES	0,1856
23	47	tucEEAFT	TEXT	AUTO	WITHOUT		0,1714

As one can see, run “tucEEAFT2” is our best run which is no surprise if one takes into account that three definitely relevant documents are used for query expansion. Our second best run is “tucEEAFTI” which shows that the additional histogram comparison produces better results than without it.

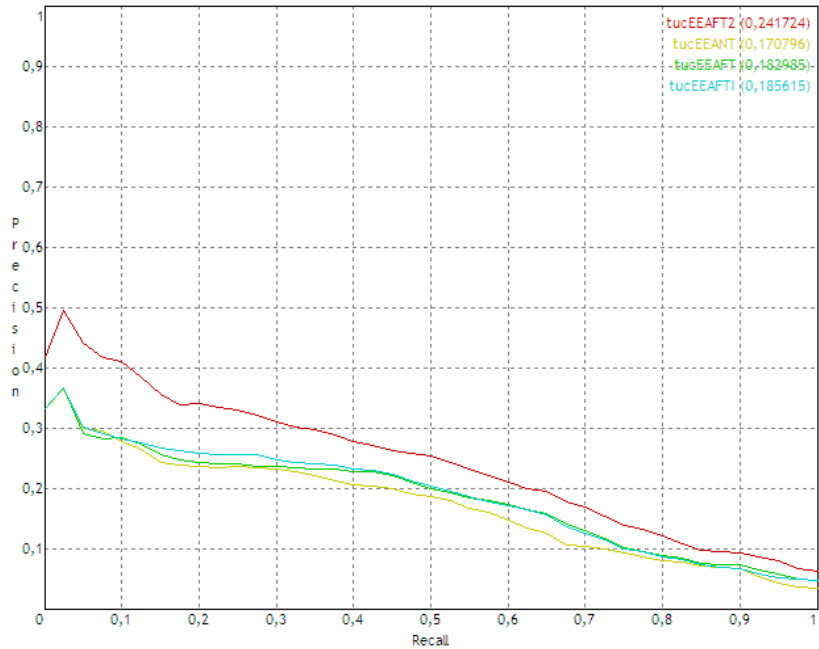


Fig. 1. Recall-precision graphs for our submitted runs

5 Conclusion

The good results in ImageCLEF 2006 show that our framework works and that our first implemented search methods provide good results.

The normal query expansion improves the mean average precision by more than 1%. Considering the poor overall mean average precision this seems to be a good

improvement of the results. Although the histogram comparison is a very simple and presumably poor measurement to compare images, it improves our text-only retrieval slightly.

The next steps to improve the results is to implement a better image retrieval algorithm due to the weak performance of the histogram comparison.

References

1. Stevens, J.S., Husted, T., Cutting, D., Carlson, P.: Apache Lucene - Overview - Apache Lucene. (2005) Retrieved August 10, 2006 from <http://lucene.apache.org/java/>
2. Grubinger, M., Leung, C., Clough, P.: The IAPR Benchmark for Assessing Image Retrieval Performance in Cross Language Evaluation Tasks (2005) Retrieved August 10, 2006 from <http://ir.shef.ac.uk/cloughie/papers/muscle-imageclef2005.pdf>
3. Roos, M.: The GNU Image-Finding Tool - GNU Project - Free Software Foundation (FSF) (2005) Retrieved August 10, 2006 from <http://www.gnu.org/software/gift/>

Overview of the CLEF-2006 Cross-Language Speech Retrieval Track

Douglas W. Oard¹, Jianqiang Wang², Gareth J.F. Jones³, Ryen W. White⁴,
Pavel Pecina⁴, Dagobert Soergel⁵, Xiaoli Huang⁵, and Izhak Shafran⁶

¹ College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
oard@umd.edu

² Department of Library and Information Studies
State University of New York at Buffalo, Buffalo, NY 14260 USA
jw254@buffalo.edu

³ School of Computing
Dublin City University, Dublin 9, Ireland
Gareth.Jones@computing.dcu.ie

⁴ Microsoft Research
One Microsoft Way, Redmond, WA 98052 USA
ryenw@microsoft.com

⁵ MFF UK, Malostranske namesti 25, Room 422
Charles University, 118 00 Praha 1, Czech Republic
pecina@ufal.mff.cuni.cz

⁶ College of Information Studies
University of Maryland, College Park, MD 20742 USA
dsoergel@umd.edu, xiaoli@umd.edu

⁷ OGI School of Science & Engineering
Oregon Health and Sciences University
20000 NW Walker Rd, Portland, OR 97006, USA
zak@cslu.ogi.edu

Abstract. The CLEF-2006 Cross-Language Speech Retrieval (CL-SR) track included two tasks: to identify topically coherent segments of English interviews in a known-boundary condition, and to identify time stamps marking the beginning of topically relevant passages in Czech interviews in an unknown-boundary condition. Five teams participated in the English evaluation, performing both monolingual and cross-language searches of speech recognition transcripts, automatically generated metadata, and manually generated metadata. Results indicate that the 2006 English evaluation topics are more challenging than those used in 2005, but that cross-language searching continued to pose no unusual challenges when compared with monolingual searches of the same collection. Three teams participated in the monolingual Czech evaluation using a new evaluation measure based on differences between system-suggested and ground truth replay start times, with results that were broadly comparable to those observed for English.

1 Introduction

The 2006 Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track continued the focus started in 2005 on ranked retrieval from spontaneous conversational speech. Automatically transcribing spontaneous speech has proven to be considerably more challenging than transcribing the speech of news anchors for the Automatic Speech Recognition (ASR) techniques on which fully-automatic content-based search systems are based.

The CLEF 2005 CL-SR task had focused solely on searching English interviews. For CLEF 2006, 30 new topics were developed for the same collection of English interviews, and an improved ASR transcript with better accuracy for the same set of interviews was added. This made it possible to validate retrieval techniques that had been shown to be effective with last year's topics, and to further explore the influence of ASR accuracy on retrieval effectiveness. The CLEF 2006 CL-SR track also added a new task of searching Czech interviews.

As with CLEF 2005, the English task was again based on a known-boundary condition for topically coherent segments. The Czech search task was based on an unknown-boundary condition in which participating systems were required to specify a replay start time for the beginning of each distinct topically relevant passage.

The first part of this paper describes the English language CL-SR task and summarizes the participants' submitted results. That is followed by a description of the Czech language task with corresponding details of submitted runs.

2 English Task

The structure of the CLEF 2006 CL-SR English task was identical to that used in 2005. Two English collections were released this year. The first release (on March 14, 2006) contained all material that was now available for training (i.e., both the training and the test topics from the CLEF 2005 CL-SR evaluation). There was one small difference from the CLEF 2005 CL-SR track data release: each person's last name that appears in the INTERVIEWDATA field (or in the associated XML data files) was reduced into its initial followed by three dots (e.g., "Smith" became "S..."). This first release contained a total of 63 topics, 8,104 topically coherent segments (the equivalent of "documents" in a classic IR evaluation), and 30,497 relevance judgments.

The second release (on June 5, 2006) included a re-release of all the training materials (unchanged), an additional 42 candidate evaluation topics (30 new topics, plus 12 other topics for which relevance judgments had not previously been released), and two new fields based on an improved ASR transcript from the IBM T. J. Watson Research Center.

2.1 Segments

Other than the changes described above, the segments used for the CLEF 2006 CL-SR task were identical to those used for CLEF 2005. Two new fields contain ASR

transcripts of higher accuracy than were available in 2005 (ASRTEXT2006A and ASRTEXT2006B). The ASRTEXT2006A field contains a transcript generated using the best presently available ASR system, which has a mean word error rate of 25% on held-out data. Because of time constraints, however, only 7,378 segments have text in this field. For the remaining 726 segments, no ASR output was available from the 2006A system at the time the collection was distributed. The ASRTEXT2006B field seeks to avoid this no-content condition by including content identical to the ASRTEXT2006A field when available, and content identical to the ASRTEXT2004A field otherwise. Since ASRTEXT2003A, ASRTEXT2004A, and ASRTEXT2006B contain ASR text that was automatically generated for all 8,104 segments, any (or all) of them can be used for the required run based on automatic data. A detailed description of the structure and fields of the English segment collection is given in last year's track overview paper [1].

2.2 Topics

A total of 30 new topics were created for this year's evaluation from actual requests received by the USC Shoah Foundation Institute for Visual History and Education.¹ These were combined with 12 topics that had been developed in previous years, but for which relevance judgments had not been released. This resulted in a set of 42 topics that were candidates for use in the evaluation.

All topics were initially prepared in English. Translations into Czech, Dutch, French, German, and Spanish were created by native speakers of those languages. With the exception of Dutch, all translations were checked for reasonableness by a second native speaker of the language.²

A total of 33 of the 42 candidate topics were used as a basis for the official 2006 CL-SR evaluation; the remaining 9 topics were excluded because they had either too few known relevant segments (fewer than 5) or too high a density of known relevant segments among the available judgments (over 48%, suggesting that many relevant segments may not have been found). Participating teams were asked to submit results for all 105 available topics (the 63 topics in the 2006 training set and the 42 topics in the 2006 evaluation candidate set) so that new pools could be formed to perform additional judgments on the development set if additional assessment resources become available.

2.3 Evaluation Measure

As in the CLEF-2005 CL-SR track, we report Mean uninterpolated Average Precision (MAP) as the principal measure of retrieval effectiveness. Version 8.0 of the `trec_eval` program was used to compute this measure.³

¹ On January 1, 2006 the University of Southern California (USC) Shoah Foundation Institute for Visual History and Education was established as the successor to the Survivors of the Shoah Visual History Foundation, which had originally assembled and manually indexed the collection used in the CLEF CL-SR track.

² A subsequent quality assurance check for Dutch revealed only a few minor problems. Both the as-run and the final corrected topics will therefore be released for Dutch.

³ The `trec_eval` program is available from http://trec.nist.gov/trec_eval/.

2.4 Relevance Judgments

Subject matter experts created multi-scale and multi-level relevance assessments in the same manner as was done for the CLEF-2005 CL-SR track [1]. These were then conflated into binary judgments using the same procedure as was used for CLEF-2005: the union of direct and indirect relevance judgments with scores of 2, 3, or 4 (on a 0–4 scale) were treated as topically relevant, and any other case as non-relevant. This resulted in a total of 28,223 binary judgments across the 33 topics, among which 2,450 (8.6%) are relevant.

2.5 Techniques

The following gives a brief description of the methods used by the participants in the English task. Additional details are available in each team's paper.

University of Alicante (UA). The University of Alicante used the MINIPAR parser to produce an analysis of syntactic dependencies in the topic descriptions and in the automatically generated portion of the collection. They then used these results in combination with their locally developed IR-n system to produce overlapping passages. Their experiments focused on combining these sources of evidence and on optimizing search effectiveness using pruning techniques.

Dublin City University (DCU). Dublin City University used two systems based on the Okapi retrieval model. One version used Okapi with their summary-based pseudo-relevance feedback method. The other system explored combination of multiple segment fields using the BM25F variant of Okapi weights. That second system was also used to explore the use of a field-based method for term selection in query expansion with pseudo-relevance feedback. The officially submitted DCU runs were adversely affected by a formatting error; the results reported for DCU in this paper are based on a subsequent re-submission that corrected that error.

University of Maryland (UMD). The University of Maryland team created two indexes, using the InQuery system in both cases. One index used only automatically generated fields, the other used only manually generated fields. Retrieval based on manually generated fields yielded MAP comparable to that reported in 2005, but retrieval based on automatically generated fields yielded MAP considerably below the 2005 values. Three CLIR experiments were conducted using French topics, but an apparent domain mismatch between the source of the translation probabilities and the CLEF CL-SR test collection resulted in relatively poor MAP for all cross-language runs.

Universidad Nacional de Educacin a Distancia (UNED). The UNED team compared the utility of the 2006 ASR with manually generated summaries

and manually assigned keywords. A CLIR experiment was performed using Spanish queries with the 2006 ASR. The system used was the same as that for their participation in the CLEF 2005 CL-SR tasks [2].

University of Ottawa (UO). The University of Ottawa used two information retrieval systems in their experiments, SMART and Terrier, each with one way of computing term weights. Two query expansion techniques were tried, including a new method based on log-likelihood scores for collocations. ASR transcripts with different estimated word error rates from 2004 and 2006 were indexed individually, together, and in combination with automatic classification results. A separate index using manually generated metadata was also built. Cross-language experiments were run using French and Spanish topics by automatically translating the topics into English using the combined results of several online machine translation tools. Results for an extensive set of locally scored runs were also reported.

University of Twente (UT). The University of Twente employed a locally developed XML retrieval system to flexibly index the collection in a way that permitted experiments to be run with different combinations of fields without reindexing. Cross-language experiments were conducted using Dutch topics, both with ASR and with manually created metadata.

2.6 English Evaluation Results

Table 2.6 summarizes the results for all 30 official runs averaged over the 33 evaluation topics, listed in descending order of MAP. Teams were asked to run at least one monolingual condition using the title and description fields of the topics and indexing only automatically generated fields; those “required runs” are shown in bold as a basis for cross-system comparisons. For the required run, DCU yielded a MAP statistically significantly better (by a Wilcoxon signed rank test for paired samples, with $p < 0.05$) than the next three teams (UO, UMD, and UT, which were statistically indistinguishable from each other). Further, the best results for this year (0.0747 from DCU) are considerably below (i.e., just 58% of) last year’s best MAP. For manually generated metadata, this year’s best result (0.2902) is 93% of last year’s best result, however, the difference is not statistically significant. From this we conclude that this year’s topic set seems somewhat less well matched with the ASR results, but that the topics are not otherwise generally much harder for information retrieval techniques based on term matching. CLIR also seemed to pose no unusual challenges with this year’s topic set, with the best CLIR on automatically generated indexing data (a French run from the University of Ottawa) achieving 83% of the MAP achieved by a comparable monolingual run. Similar effects were observed with manually generated metadata (at 80% of the corresponding monolingual MAP for Dutch queries, from the University of Twente).

Table 1. English official runs. Bold runs are the required condition. N = Name (Manual metadata), MK = Manual Keywords (Manual metadata), SUM = Summary (Manual metadata), ASR4 = ASRTEXT2004A (Automatic), ASR6 = ASRTEXT2006B (Automatic), AK1 = AUTOKEYWORD2004A1 (Automatic), AK2 = AUTOKEYWORD2004A2. See [1] for descriptions of these fields. (Automatic).

Run name	MAP	Lang	Query	Doc field	Site
uoEnTDNtMan	0.2902	EN	TDN	MK,SUM	UO
3d20t40f6sta5flds	0.2765	EN	TDN	ASR6,AK1,AK2,N,SUM,MK	DCU
umd.manu	0.2350	EN	TD	N,MK,SUM	UMD
UTsummkeNnor	0.2058	EN	T	MK,SUM	UT
dcuEgTDall	0.2015	EN	TD	ASR6,AK1,AK2,N,SUM,MK	DCU
uneden-manualkw	0.1766	EN	TD	MK	UNED
UTsummknI2or	0.1654	NL	T	MK,SUM	UT
dcuFchTDall	0.1598	FR	TD	ASR6,AK1,AK2,N,SUM,MK	DCU
umd.manu.fr.0.9	0.1026	FR	TD	N,MK,SUM	UMD
umd.manu.fr.0	0.0956	FR	TD	N,MK,SUM	UMD
unedes-manualkw	0.0904	ES	TD	MK	UNED
unedes-summary	0.0871	ES	TD	SUM	UNED
uoEnTDNsQEx04A	0.0768	EN	TDN	ASR4,AK1,AK2	UO
dcuEgTDauto	0.0733	EN	TD	ASR6,AK1,AK2	DCU
uoFrTDNs	0.0637	FR	TDN	ASR4,AK1,AK2	UO
uoSpTDNs	0.0619	ES	TDN	ASR4,AK1,AK2	UO
uoEnTDt04A06A	0.0565	EN	TD	ASR4,ASR6,AK1,AK2	UO
umd.auto	0.0543	EN	TD	ASR4,ASR6,AK1,AK2	UMD
UTasr04aEN	0.0495	EN	T	ASR4	UT
dcuFchTDauto	0.0462	FR	TD	ASR6,AK1,AK2	DCU
UA_TDN_FL_ASR06BA1A2	0.0411	EN	TDN	ASR6,AK1,AK2	UA
UA_TDN_ASR06BA1A2	0.0406	EN	TDN	ASR6,AK1,AK2	UA
UA_TDN_ASR06BA2	0.0381	EN	TDN	ASR6,AK2	UA
UTasr04aNI2	0.0381	NL	T	ASR4	UT
UTasr04aEN-TD	0.0381	EN	TD	ASR4	UT
uneden	0.0376	EN	TD	ASR6	UNED
UA_TD_ASR06B	0.0375	EN	TD	ASR6	UA
UA_TD_ASR06BA2	0.0365	EN	TD	ASR6,AK2	UA
unedes	0.0257	ES	TD	ASR6	UNED
umd.auto.fr.0.9	0.0209	FR	TD	ASR4,ASR6,AK1,AK2	UMD

3 Czech Task

The goal of the Czech task was to automatically identify the start points of topically-relevant passages in unsegmented interviews. Ranked lists for each topic were submitted by each system in the same form as the English task, with the single exception that a system-generated starting point was specified rather than a document identifier. The format for this was “VHF[IntCode].[starting-time],” where “IntCode” is the five-digit interview code (with leading zeroes added) and “starting-time” is the system-suggested replay starting point (in seconds) with

reference to the beginning of the interview.⁴ Lists were to be ranked by systems in the order that the system would suggest for listening to passages beginning at the indicated points.

3.1 Interviews

The Czech task was broadly similar to the English task in that the goal was to design systems that could help searchers identify sections of an interview that they might wish to listen to. The processing of the Czech interviews was, however, different from that used for English in three important ways:

- No manual segmentation was performed. This alters the format of the interviews (which for Czech is time-oriented rather than segment-oriented), it alters the nature of the task (which for Czech is to identify replay start points rather than to select among predefined segments), and it alters the nature of the manually assigned metadata (there are no manually written summaries for Czech and the assignment of a manual thesaurus term implies that the discussion of a topic started at that time).
- The two available Czech ASR transcripts (2004 and 2006) were generated using different ASR systems. In both cases, the acoustic models were trained using 15-minutes snippets from 336 speakers, all of whom are present in the test set as well. However, the language model for the 2006 system was created by interpolating two models—an in-domain model from transcripts, and an out-of-domain model from selected portions of the Czech National Corpus. For details, see the baseline systems described in [3,4]. Apart from the improvement in transcription accuracy, the 2006 system differs from the 2004 system in that the transcripts are produced in formal Czech, rather than the colloquial Czech that was produced in 2004. Since the topics were written in formal Czech, the 2006 ASR transcripts would be expected to yield better matching. Interview-specific vocabulary priming (adding proper names to the recognizer vocabulary based on names present in a pre-interview questionnaire) was not done for either Czech system. Thus, a somewhat higher error rate on named entities might be expected for the Czech systems than for the two English systems (2004 and 2006) in which vocabulary priming was included.
- ASR is available for both the left and right stereo channels (which were usually recorded from microphones with different positions and orientations).

Because the task design for Czech is not directly compatible with the design of document-oriented IR systems, we provided a “quickstart” package containing the following:

⁴ The interviews were initially recorded on separate tapes, and we adopted the convention that the start time of a tape was defined to be the end time of the last word recognized by the 2006 system for the channel that yielded the largest number of recognized words (across all tapes) for that system. Note that these tape start times are the same for all four transcripts (2004 and 2006, left and right channels).

- A quickstart script for generating overlapping passages directly from the ASR transcripts. The passage duration (in seconds), the spacing between passage start times (also in seconds), and the desired ASR system (2004 or 2006) could be specified. The default settings (180, 60, and 2006) were intended to result in 3-minute passages with start times spaced one minute apart. A design error in the quickstart scripts resulted in use of the sum of the word durations rather than the word start times; hence the default passage length turned out to be about 4 minutes (and, somewhat more precisely, an average of 403 words).
- A quickstart collection created by running the quickstart script with the default settings. This collection contains 11,377 overlapping passages.

The quickstart collection contains the following automatically generated fields:

DOCNO. The DOCNO field contains a unique document number in the same format as the start times that systems were required to produce in a ranked list. This design allowed the output of a typical IR system to be used directly as a list of correctly formatted (although perhaps not very accurate) start times for scoring purposes.⁵

ASRSYSTEM. specifying the type of the ASR transcript, where “2004” and “2006” denote colloquial and formal Czech transcripts respectively. In addition, the “2006” transcript benefited from improvements in the ASR system, which were developed at Johns Hopkins University with input from University of West Bohemia. By default, the quickstart collection would use 2006 transcripts when they were available; otherwise 2004 transcripts were used.⁶

CHANNEL. The CHANNEL field specifies which recorded channel (left or right) was used to produce the transcript. The channel that produced the greatest number of total words over the entire transcript (which is usually the channel that produced the best ASR accuracy for words spoken by the interviewee) was automatically selected by default. This automatic selection process was hardcoded in the script, although the script could be modified to generate either or both channels.

ASRTEXT. The ASRTEXT field contains words in order from the transcript selected by ASRSYSTEM and CHANNEL for a passage beginning at the start time indicated in DOCNO. When the selected transcript contains no words at all for a tape, words are drawn from one alternate source that is chosen in the following priority order: (1) the same ASRSYSTEM from the other CHANNEL, (2) the same CHANNEL from the other ASRSYSTEM, or (3) the other CHANNEL from the other ASRSYSTEM.

⁵ The design error in the script that generated the quickstart collection resulted in incorrect times appearing in the DOCNO field in the distributed collection. Results reported in this volume are based on automatic correction of the times in each DOCNO.

⁶ Transcripts for some tapes were not available from the 2006 Czech ASR system at the time the collection was built.

ENGLISHAUTOKEYWORD. The ENGLISHAUTOKEYWORD field contains a set of thesaurus terms that were assigned automatically using a k-Nearest Neighbor (kNN) classifier based solely on words from the ASRTEXT field of the passage; the top 20 thesaurus terms are included in best-first order. Thesaurus terms (which may be phrases) are separated with a vertical bar character. The classifier was trained using English data (manually assigned thesaurus terms and manually written segment summaries) and run on automatically produced English translations of the 2006 Czech ASRTEXT [5]. Two types of thesaurus terms are present, but not distinguished: (1) terms that express a subject or concept; (2) terms that express a location, often combined with time in one precombined term [6]. Because the classifier was trained on the English collection, in which thesaurus terms were assigned to segments, the natural interpretation of an automatically assigned thesaurus term is that the classifier believes the indicated topic is associated with the words spoken in this passage. Note that this differs from the way in which the presence of a manually assigned thesaurus term (described below) should be interpreted.

CZECHAUTOKEYWORD. The CZECHAUTOKEYWORD field contains Czech translations of the ENGLISHAUTOKEYWORD field. These translations were obtained from three sources: (1) professional translation of about 3,000 thesaurus terms, (2) volunteer translation of about 700 thesaurus terms, and (3) a custom-built machine translation system that reused words and phrases from manually translated thesaurus terms to produce additional translations [6]. Some words (e.g., foreign place names) remained untranslated when none of the three sources yielded a usable translation.

Three additional fields containing data produced by human indexers at the Survivors of the Shoah Visual History Foundation were also available for use in contrastive conditions:

INTERVIEWDATA. The INTERVIEWDATA field contains the first name and last initial for the person being interviewed. This field is identical for every passage that was generated from the same interview.

ENGLISHMANUKEYWORD. The ENGLISHMANUALKEYWORD field was intended to contain thesaurus terms that were manually assigned with one-minute granularity from a custom-built thesaurus by subject matter experts at the Survivors of the Shoah Visual History Foundation while viewing the interview. The format is the same as that described for the ENGLISHAUTOKEYWORD field, but the meaning of a keyword assignment is different. In the Czech collection, manually assigned thesaurus terms are used as on-set marks—they appear only once at the point where the indexer recognized that a discussion of a topic or location-time pair had started; continuation and completion of discussion are not marked. Unfortunately, the design error in the quickstart script resulted in misplacement of the manually assigned thesaurus terms in later segments of the quickstart collection than was appropriate.

CZECHMANUKEYWORD. The CZECHMANUKEYWORD field contains Czech translations of the English thesaurus terms that were produced from the ENGLISHMANUKEYWORD field using the process described above.⁷

All three teams used the quickstart collection; no other approaches to segmentation and no other settings for passage length or passage start time spacing were tried.

3.2 Topics

At the time the Czech evaluation topics were released, it was not yet clear which of the available topics were likely to yield a sufficient number of relevant passages in the Czech collection. Participating teams were therefore asked to run 115 topics—every available topic at that time. This included the full 105 topic set that was available in 2006 for English (including all training and all evaluation candidate topics) and adaptations of 10 topics from that set in which geographic restrictions had been removed (as insurance against the possibility that the smaller Czech collection might not have adequate coverage for exactly the same topics).

All 115 topics had originally been constructed in English and then translated into Czech by native speakers. Since translations into languages other than Czech were not available for the 10 adapted topics, only English and Czech topics were distributed with the Czech collection. No teams used the English topics this year; all official runs this year with the Czech collection were monolingual.

Two additional topics were created as part of the process of training relevance assessors, and those topics were distributed to participants along with a (possibly incomplete) set of relevance judgments. However, this distribution occurred too late to influence the design of any participating system.

3.3 Evaluation Measure

The evaluation measure that we chose for Czech is designed to be sensitive to errors in the start time, but not in the end time, of system-recommended passages. It is computed in the same manner as mean average precision, but with one important difference: partial credit is awarded in a way that rewards system-recommended start times that are close to those chosen by assessors. After a simulation study, we chose a symmetric linear penalty function that reduces the credit for a match by 0.1 (absolute) for every 15 seconds of mismatch (either early or late) [7]. This results in the same computation as the well-known mean Generalized Average Precision (mGAP) measure that was introduced to deal with human assessments of partial relevance [8]. In our case, the human assessments are binary; it is the degree of match to those assessments that can be partial. Relevance judgments are used without replacement, which means that only the highest

⁷ Because the ENGLISHMANUKEYWORD field is not correct in the quickstart collection, the CZECHMANUKEYWORD field is also incorrect.

ranked match (including partial matches) can be scored for any relevance assessment; other potential matches receive a score of zero. Differences at or beyond a 150 second error are treated as a no-match condition, thus not “using up” a relevance assessment. Systems could therefore optimize their mGAP score by automatically removing near-duplicates (i.e., start points that are close in time) that occur lower in the list, although no system did so this year.

3.4 Relevance Judgments

Relevance judgments were completed at Charles University in Prague for a total of 29 Czech topics by subject matter experts who were native speakers of Czech. All relevance assessors had good English reading skills. Topic selection was performed by individual assessors, subject to the following factors:

- At least five relevant start times in the Czech collection were required in order to minimize the effect of quantization noise on the computation of mGAP, so assessors were encouraged to perform initial triage searches and then focus on topics that they expected to meet this criterion.
- The greatest practical degree of overlap with topics for which relevance judgments were available in the English collection was desirable, so choosing to assess one of the “adapted” topics (for which no English relevance judgments are available) was discouraged.

Once a topic was selected, the assessor iterated between topic research (using external resources) and searching the collection. A new search system was designed to support this interactive search process. The best channel of the Czech ASR and the manually assigned English thesaurus terms were indexed as overlapping passages, and queries could be formed using either or both. Once a promising interview was found, an interactive search within the interview could be performed using either type of term. Promising regions (according to the system) were then highlighted for the assessor using a graphical depiction of the retrieval status value. Assessors could then scroll through the interview using that graphical depiction in conjunction with the displayed English thesaurus terms and the displayed ASR transcript as a basis for recognizing regions that they believed might be topically relevant. They could then replay the audio from any point in order to confirm topical relevance. As they did this, they could indicate the onset and conclusion of the relevant period by designating points on the transcript that were then automatically converted to times with 15-second granularity.⁸ Only the start times are used for computation of the mGAP measure, but both start and end times are available for future research.

⁸ Several different types of time spans arise when describing evaluation of speech indexing systems. For clarity, we have tried to stick to the following terms when appropriate: manually defined segments (for English indexing), 15-minute snippets (for ASR training), 15-second increments (for the start and end time of Czech relevance judgments), relevant passages (identified by Czech relevance assessors), and automatically generated passages (for the quickstart collection).

Once that search-guided relevance assessment process was completed, the assessors were provided with a set of additional “highly ranked” points to check for topical relevance that were computed using a pooling technique similar to that used for English. The top 50 start times from every official run were pooled, duplicates (at one minute granularity) were removed, and the results were inserted into the assessment system as system recommendations. Every system recommendation was checked, although assessors exercised judgment regarding when it would be worthwhile to actually listen to the audio in order to limit the cost of this process. Relevant passages identified in this way were added to those found using search-guided assessment to produce the final set of relevance judgments (topic 4000 was generalized from a pre-existing topic).⁹

Table 2. Number of the relevant passages identified for each of the 29 topics in the Czech collection

topic	#rel	topic	#rel	topic	#rel	topic	#rel	topic	#rel
1166	8	1181	21	1185	50	1187	26	1225	9
1286	70	1288	9	1310	20	1311	27	1321	27
14312	14	1508	83	1620	35	1630	17	1663	34
1843	52	2198	18	2253	124	3004	43	3005	84
3009	77	3014	87	3015	50	3017	83	3018	26
3020	67	3025	51	3033	45	4000	65		

A total of 1,322 start times for relevant passages were identified, thus yielding an average of 46 relevant passages per topic (minimum 8, maximum 124). Table 2 shows the number of relevant start times for each of the 29 topics, 28 of which are the same as topics used in the English test collection (the exception being topic 4000, an “adapted” topic).

3.5 Techniques

The participating teams all employed existing information retrieval systems to perform monolingual searches of the quickstart collection.

University of Maryland (UMD). The University of Maryland submitted three runs in which they tried combining all the fields (the Czech ASR transcript, automatically and manually assigned Czech thesaurus terms, and the English translations of those thesaurus terms) to form a unified passage index again using Inquiry. They compared the retrieval results based on that index with results based on indexing ASR alone or based on indexing only a combination of automatically assigned thesaurus terms and the ASR transcript.

⁹ Because all three participating teams used the quickstart collection, which had incorrect segment start times, this highly ranked assessment process was likely less effective than it should have been. For purposes of characterizing the relevance judgments for the 2006 Czech collection, it is therefore reasonable to consider them as having been created principally using a search-guided assessment process.

University of Ottawa (UO). Four runs were submitted from the University of Ottawa using SMART, and a fifth run was submitted using Terrier. The configuration of the SMART runs explored indexed field combinations similar to those tried by the University of Maryland; the Terrier run used ASR in combination with automatically assigned thesaurus terms.

University of West Bohemia (UWB). The University of West Bohemia was the only team to apply morphological normalization and stopword removal for Czech. A TF*IDF model was implemented in Lemur, along with the Lemur implementation of blind relevance feedback. Five runs were submitted for official scoring, and additional runs were scored locally.

Results. The results reported in Table 3 were computed after correction of the times in the DOCID field of the quickstart collection. Because this correction invalidates the use of manually assigned thesaurus terms, their use (and use of their translations) should be interpreted only as a source of noise. As a comparison of runs UWB_mk_aTD and UWB_mk_a_akTD shows, this effect is small. Somewhat surprisingly, there is no evidence that automatically assigned English thesaurus terms or their Czech translations was helpful. Examination of a sample of translation pairs by a native speaker identified remarkably few translation errors, so it seems reasonable to dismiss that factor as a possible cause. The automatic assignment process was based on the Czech terms from the same segment, so time misalignment also seems unlikely to be the cause. Manual examination of a sample of the English thesaurus terms does however, indicate that the assignments appear to bear little relation to the topical content of the segment. We therefore suspect either a deficiency in the classifier design or some broader-scale alignment error in associating the classifier results with ASR segments. That issue has not yet been resolved, and for the analysis in this paper we simply treat the use of automatically assigned thesaurus terms (in either language) as an additional source of noise.

Every run except uoCzEnTDNsMan used the Czech ASR transcript, so we consider those 12 official runs as a group for the purpose of this analysis (ignoring the use of other fields). A strict ordering by participating site is evident, with UWB achieving better results with every one of their runs than any run from any other site. This suggests that handling Czech morphology in some way is particularly important.

These results can almost certainly be improved upon in future experiments. We had originally intended the quickstart collection to be used only for out-of-the-box sanity checks, with the idea that teams would either modify the quickstart scripts or create new systems outright to explore a broader range of possible system designs. Time pressure and a lack of a suitable training collection precluded that sort of experimentation, however, and the result was that this undesirable effect of passage overlap affected every system. The use of relatively long overlapping passages in the quickstart collection probably reduced

Table 3. Czech official runs with passage start times corrected to match ASR. ASR = ASRTEXT, CAK = CZECHAUTOKEYWORD (Automatic), EAK = ENGLISHAUTOKEYWORD (Automatic), CMK = CZECHMANUKEYWORD (Manual metadata), EMK = ENGLISHMANUKEYWORD (Manual metadata).

Run name	mGAP	Lang	Query	Doc field	Site
UWB_mk_a_akTDN	0.0456	CZ	TDN	ASR,CAK,CMK	UWB
UWB_aTD	0.0435	CZ	TD	ASR	UWB
UWB_mk_aTD	0.0416	CZ	TD	ASR,CMK	UWB
UWB_a_akTD	0.0402	CZ	TD	ASR,CAK	UWB
UWB_mk_a_akTD	0.0377	CZ	TD	ASR,CAK,CMK	UWB
uoCzEnTDNsMan	0.0235	CZ,EN	TDN	CAK,CMK,EAK,EMK	UO
uoCzEnTDt	0.0218	CZ,EN	TD	ASR,CAK	UO
uoCzTDs	0.0211	CZ	TD	ASR,CAK	UO
uoCzTDNsMan	0.0200	CZ	TDN	ASR,CAK,CMK	UO
uoCzTDNs	0.0182	CZ	TDN	ASR,CAK	UO
umd.all	0.0070	CZ	TD	ASR,CAK,CMK,EAK,EMK	UMD
umd.asr	0.0066	CZ	TD	ASR	UMD
umd.akey.asr	0.0060	CZ	TD	ASR,CAK,EAK	UMD

mGAP values somewhat because the (roughly) 80-second passage spacing was nearly five times the 15-second granularity of the relevance judgments.

4 Conclusion and Future Plans

The CLEF 2006 CL-SR track extended the previous year's work on the English task by adding new topics, and by introducing a new Czech task with a new unknown-boundary evaluation condition. The results of the English task suggest that the evaluation topics this year posed somewhat greater difficulty for systems doing fully automatic indexing. Studying what made these topics more difficult would be an interesting focus for future work. However, the most significant achievement of this year's track was the development of a CL-SR test collection based on a more realistic unknown-boundary condition. Now that we have both that collection and an initial set of system designs, we are in a good position to explore issues of system and evaluation design.

The CLEF CL-SR track will continue in 2007. For Czech, relevance judgments will be created for additional topics, with a goal of having at least 50 topic for Czech as a legacy from the track for use by future researchers. The present set of 96 topics for English is already adequate for many future purposes, and indeed for a collection of this size construction of additional topics that are representative of real information needs could prove to be impractical. We therefore will run the English task again in 2007 with the same evaluation topics in order to support comparative evaluation of improved systems, but no new topics are planned. Some unique characteristics of the CL-SR collection (e.g., location-oriented topics) may also be of interest to other tracks, including

domain-specific retrieval and geoCLEF. By sharing our resources widely, we hope to maximize the impact of the pioneering research teams that have contributed to the construction of these unique resources.

Acknowledgments

This track would not have been possible without the efforts of a great many people. Our heartfelt thanks go to the dedicated group of relevance assessors in Maryland and Prague, to the Dutch, French and Spanish teams that helped with topic translation, and to Bill Byrne, Martin Cetkovsky, Bonnie Dorr, Ayelet Goldin, Sam Gustman, Jan Hajic, Jimmy Lin, Baolong Liu, Craig Murray, Scott Olsson, Bhuvana Ramabhadran and Deborah Wallace for their help with creating the techniques, software, and data sets on which we have relied.

References

1. White, R.W., Oard, D.W., Jones, G.J.F., Soergel, D., Huang, X.: Overview of the clef-2005 cross-language speech retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. López-Ostenero, F., Peinado, V., Sama, V., Verdejo, F.: UNED@CL-SR CLEF 2005: Mixing different strategies to retrieve automatic speech transcriptions. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
3. Shafran, I., Byrne, W.: Task-specific minimum bayes-risk decoding using learned edit distance. In: Proceedings of INTERSPEECH2004-ICSLP (2004)
4. Shafran, I., Hall, K.: Corrective models for speech recognition of inflected languages. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2006)
5. Olsson, J.S., Oard, D.W., Hajic, J.: Cross-language text classification. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York (2005)
6. Murray, G.C., Dorr, B.J., Lin, J., Hajic, J., Pecina, P.: Leveraging reusability: Cost-effective lexical acquisition for large-scale ontology translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (2006)
7. Liu, B., Oard, D.W.: One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York (2006)
8. Kekalainen, J., Jarvelin, K.: Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology* (2002)

Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006*

Pavel Ircing and Luděk Müller

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{ircing, muller}@kky.zcu.cz

Abstract. The paper describes the system built by the team from the University of West Bohemia for participation in the CLEF 2006 CL-SR track. We have decided to concentrate only on the monolingual searching in the Czech test collection and investigate the effect of proper language processing on the retrieval performance. We have employed the Czech morphological analyser and tagger for that purposes. For the actual search system, we have used the classical *tf.idf* approach with blind relevance feedback as implemented in the Lemur toolkit. The results indicate that a suitable linguistic preprocessing is indeed crucial for the Czech IR performance.

1 Introduction

This paper presents the first participation of the University of West Bohemia group in CLEF (and, for that matter, first participation of the group in an IR evaluation campaign whatsoever). Thus, being novices in the IR field, we have decided to concentrate only on the monolingual searching in the Czech test collection where we have tried to make use of the two advantages that our team might have over the others — the knowledge of the language in question (Czech — our mother tongue) and the experience with automatic NLP of that language, together with the employment of the necessary tools (morphological analyser, tagger).

As for the actual search side of the task, it has been shown by various teams experimenting with last year's English test collection that good results can be achieved simply by using some freely available IR system (see for example [1]). We have decided to use the same strategy.

Although both the English and the Czech CL-SR collections consists of the (automatic) transcriptions of the interviews with the Holocaust survivors (plus some additional metadata — see the description of the collections in the track

* This work was supported by the Grant Agency of the Czech Academy of Sciences project No. 1ET101470416 and the Ministry of Education of the Czech Republic project No. LC536.

overview [2]), the Czech collection lacks the manually created topical segmentation that is available for the English data. This obviously makes the retrieval more complicated. Thus, in order to facilitate the initial experiments with the Czech collection, the track organizers provided also a so-called Quickstart collection with artificially defined “documents” that were created by sliding 3-minute window over the continuous stream of transcriptions with the 1-minute step. Given the lack of time for experimentation, the presence of many other system parameters and the absence of training topics, we did not explore any other segmentation possibilities beyond this Quickstart collection in our experiments. It turned out at the workshop that all other teams applied the same approach which makes all the results well comparable.

2 System Description

2.1 Linguistic Preprocessing

At least rudimentary linguistic processing of the document collection and topics (stemming, stop-word removal) is considered to be indispensable in state-of-the-art IR systems. We have decided to use quite sophisticated NLP tools for that purpose — the morphological analyser and tagger developed by the team around Jan Hajič [3],[4]. The serial combination of these two tools assigns disambiguated lemma (basic word form) and morphological tag to the input word form and also provides the information about the stem-ending partitioning.

This is an example of the typical system output:

```
<f>holokaustem<MD1>holokaust<MDt>NNIS7-----A----<R>holokaust<E>em
```

where <f> introduces the actual word form, the <MD1> the corresponding lemma and the <MDt> the corresponding morphological tag (in this case the tag correctly describes the word **holokaustem** as the noun (N in the first position) having the masculine inanimate gender (I) and being in singular (S) instrumental (7) form). Finally, the <R> introduces the stem and <E> the ending of the word form in question. Note that although in this example the stem is identical to the lemma it is not the general rule. Thus one of the issues to be resolved is whether to use lemmatization or stemming for the data processing. For English, the results typically favor neither lemmatization nor stemming over each other [5] and since the stemming is simpler, it is most commonly used. As our tools perform Czech lemmatization and stemming in a single step, we investigated both strategies in our experiments.

The information provided by the NLP tools was also exploited for stop-word removal. As we were not able to find any decent stoplist of Czech words we have decided to remove words on the basis of their part-of-speech (POS). As can be seen from the example above, the POS information is present at the first position of the morphological tag. We removed from indexing all the words that were tagged as prepositions, conjunctions, particles and interjections (note that they are no articles in Czech).

Here is an example of one of the topics before and after the linguistic pre-processing

The original topic:

```
<top>
<num>1286</num>
<title>Hudba v holokaustu</title>
<desc>Svědectví o tom, zda hudba pomáhala (duševně nebo i jinak)
nebo překážela vězňům internovaným v koncentračních táborech.
</desc>
<narr>Popis toho, jakou roli hrála hudba v životě vězňů.</narr>
</top>
```

gets processed into:

```
<top>
<num>1286</num>
<title>hudba holokaust</title>
<desc>svědectví ten hudba pomáhat duševně jinak překážet vězeň
internovaný koncentrační tábor</desc>
<narr>popis ten jaký role hrát hudba život vězeň</narr> </top>
```

when using lemmatization or into:

```
<top>
<num>1286</num>
<title>hud holokaust</title>
<desc>svědectví tom hud pomáh duševně jinak překáž vězňům
internovan koncentračn tábor</desc>
<narr>popis toho jakou rol hrál hud život vězňů</narr> </top>
```

when the stemming is employed.

2.2 Retrieval

For the actual IR we have used the freely available Lemur toolkit [6] that allows us to employ various retrieval strategies, including among others the classical vector space model and the language modeling approach.

We have decided to stick to the *tf.idf* model where both documents and queries are represented as weighted term vectors $\mathbf{d}_i = (w_{i,1}, w_{i,2}, \dots, w_{i,n})$ and $\mathbf{q}_k = (w_{k,1}, w_{k,2}, \dots, w_{k,n})$, respectively (n denotes the total number of distinct terms in the collection). The inner-product of such weighted term vectors then determines the similarity between individual documents and queries. As there are many ways to compute the weights $w_{i,j}$ without any of them performing consistently better than the others, we employed the very basic formula

$$w_{i,j} = tf_{i,j} \cdot \log \frac{d}{df_j} \quad (1)$$

where $tf_{i,j}$ denotes the number of occurrences of the term t_j in the document d_i (term frequency), d is the total number of documents in the collection and finally df_j denotes the number of documents that contain t_j . We have not used any document length normalization as the length of the “documents” in the Quickstart collection is approximately uniform.

In order to boost the performance, we also used the simplified version of the blind relevance feedback implemented in Lemur [7]. The original Rocchio’s algorithm is defined by the formula

$$\mathbf{q}_{new} = \mathbf{q}_{old} + \alpha \cdot \mathbf{d}_R - \beta \cdot \mathbf{d}_{\bar{R}} \quad (2)$$

where R and $\bar{R}$ denote the set of relevant and non-relevant documents, respectively, and $\mathbf{d}_R$ and $\mathbf{d}_{\bar{R}}$ denote the corresponding centroid vectors of those sets. In other words, the basic idea behind this algorithm is to move the query vector closer to the relevant documents and away from the non-relevant ones. In the case of blind feedback, the top M documents from the first-pass run are simply treated as if they were relevant. The Lemur modification of this algorithm sets the $\beta = 0$ and keeps only the K top-weighted terms in $\mathbf{d}_R$.

3 Experimental Evaluation

As we already mentioned in the Introduction, all the experiments were carried out on the Czech Quickstart collection, using only the Czech version of the queries. There are 115 queries defined for searching in the Czech test collection. However, only 29 of them were manually evaluated by the assessors and used to generate the `qrel` files.

3.1 Evaluated Runs

Table 1 summarizes the results for the runs that we consider to be important. The mean Generalized Average Precision (mGAP) is used as the evaluation metric — the details about this measure can be found in [8]. The table contains two main horizontal sections — the upper one contains the results achieved when using only `<title>` and `<desc>` fields of the topics as queries (TD), the lower one then the results with all `<title>`, `<desc>` and `<narr>` fields used (TDN). Inside each of these sections, the runs are further divided according to the type of the performed linguistic processing — both the queries and the collection were always stoplisted and either left in the original word forms (w) or stemmed (s) or lemmatized (l). Finally, the vertical division denotes the fields from the collection that were indexed for the corresponding runs — `<CZECHMANUKEYWORD>` (mk), `<ASRTEXT>` (asr), `<CZECHAUTOKEYWORD>` (ak) and their combinations. The runs typeset in boldface are the ones that were submitted for the official CLEF scoring.

Table 1. Mean GAP of the individual runs - bold runs were submitted for official scoring

		mk	asr	ak	mk.asr	asr.ak	mk.asr.ak
TD	w	0.0026	0.0271	0.0022	0.0270	0.0240	0.0247
	s	0.0030	0.0438	0.0024	0.0441	0.0405	0.0401
	l	0.0026	0.0435	0.0024	0.0416	0.0402	0.0377
TDN	w	0.0025	0.0256	0.0018	0.0266	0.0241	0.0242
	s	0.0029	0.0494	0.0022	0.0488	0.0447	0.0456
	l	0.0024	0.0506	0.0023	0.0518	0.0467	0.0456

3.2 Analysis of the Results

First, we cannot resist to mention that all our submitted runs significantly outperformed the runs of the other two Czech sub-track participants. The reason of this is quite simple and readily apparent from the table of results — as far as we know, we were the only team that used Czech lemmatization and/or stemming and the table shows that either one of these operations boosts the IR performance almost by a factor of two (that is, at least in the columns where the asr field is indexed).

What the experiments did not help to resolve is whether to use stemming or lemmatization in the preprocessing stage — both methods yielded comparable results. On the other hand, the results show that both the manual and the automatic keyword fields are virtually useless for the retrieval — they perform poorly when indexed alone and do not bring any noticeable improvement when added to the asr field. Further investigation revealed that the scripts that were used for the creation of the Quickstart collection contain systematic error that caused manual keywords to be assigned to the wrong segments. Manual examination of a sample of the automatically assigned English thesaurus terms indicates that the automatic keyword assignment is not overly accurate either. Therefore both the manual and the automatic keywords most probably bring more noise than useful information to the IR system.

We have also performed a couple of experiments in order to investigate the effect of the stop-word removal (not shown). We have found out that removing the stop words on the basis of morphological tags did not help the IR performance in comparison with working with the “non-stopped” collection and queries. On the other hand, it substantially reduced the size of the index files.

3.3 Tips for Future Improvement

First of all, let us point out again that the segments in the Quickstart collections are not well-formed “documents”, especially in the sense that they are not, in most cases, topically coherent. Thus a more sophisticated way of collection segmentation might be useful. Moreover, the quality of the ASR transcriptions is rather poor — around 35% WER in general but even more for the named entities (NEs) which are extremely important for searching. The application of

the specially designed language model focused on a more accurate transcription of NEs could partially alleviate this problem [9]. Finally, there appears to be a non-negligible vocabulary mismatch between the topics and the collection or even between the different fields in the collection. For example, just looking at the first two topics that were evaluated by the assessors we have discovered that in topic 1181 the name of the infamous concentration camp “Auschwitz” was kept untranslated in the topics but it was translated into its Czech form (“Osvětim”) in the <CZECHMANUKEYWORD> and <CZECHAUTOKEYWORD> fields,¹ the word “Sonderkommando” was written with double “m” in the topics and in the <ASRTEXT> field and with single “m” in the keyword fields. A coherent treatment of such different variants is therefore desirable.

4 Conclusion

The Czech CL-SR track presents a first attempt to create and test the collection of the Czech spontaneous speech. As such, it suffered some initial difficulties that were for the most part identified by the end of the actual CLEF workshop and just recently fixed. As a result, we are currently at the performance level comparable with the English part of the track but the lack of time prevented us from doing even more detailed result analysis (especially significance testing). Nevertheless, we are already in a good position for the next year’s campaign where the current evaluation set of topics will most probably serve as a training data.

References

1. Inkpen, D., Alzghool, M., Islam, A.: Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Mueller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 760–768. Springer, Heidelberg (2006)
2. Oard, D., Wang, J., Jones, G., White, R., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In: Peters, C., Clough, P., Gey, F., Karlgren, J., Magnini, B., Oard, D., de Rijke, M., Stempfhuber, M. (eds.) Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain. LNCS (to appear, 2007)
3. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Prague (2004)
4. Hajič, J., Hladká, B.: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings of COLING-ACL Conference, Montreal, Canada, pp. 483–490 (1998)
5. Hull, D.A.: Stemming Algorithms: A Case Study for Detailed Evaluation. Journal of the American Society of Information Science 47(1), 70–84 (1996)

¹ Note that neither one of those variants is the original name of the Polish town in question — “Oświęcim.” Consequently, all three forms are routinely used by the Czech speakers and therefore appear in the ASR transcripts.

6. Carnegie Mellon University and the University of Massachusetts: The Lemur Toolkit for Language Modeling and Information Retrieval (2006), <http://www.lemurproject.org/>
7. Zhai, C.: Notes on the Lemur TFIDF model. Note with Lemur 1.9 documentation, School of CS, CMU (2001)
8. Liu, B., Oard, D.: One-Sided Measures for Evaluating Ranked Retrieval Effectiveness with Spontaneous Conversational Speech. In: Proceedings of SIGIR 2006, Seattle, Washington, USA, pp. 673–674 (2006)
9. Ircing, P., Psutka, J., Radová, V.: Automatic Transcription of Audio Archives for Spoken Document Retrieval. In: Proceedings of Computational Intelligence 2006, San Francisco, USA, pp. 448–452 (2006)

Applying Logic Forms and Statistical Methods to CL-SR Performance^{*}

Rafael M. Terol, Patricio Martinez-Barco, and Manuel Palomar

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera de San Vicente del Raspeig - Alicante - Spain
Tel.: +34965903772 Fax: +34965909326
{rafamt, patricio, mpalomar}@dlsi.ua.es

Abstract. This paper describes a CL-SR system that employs two different techniques: the first one is based on NLP rules that consist on applying logic forms to the topic processing while the second one basically consists on applying the IR-n statistical search engine to the spoken document collection. The application of logic forms to the topics allows to increase the weight of topic terms according to a set of syntactic rules. Thus, the weights of the topic terms are used by IR-n system in the information retrieval process.

1 Introduction

Continuing with the previous research activities developed by the University of Alicante that were presented in previous CLEF CL-SR track, our system runs according to two different stages: the first one consists on increasing the weight of some topic terms by applying a set of rules that are based on the representation of the topics by means of logic forms [3]; and the second one consists of only applying IR-n system [1] to the spoken document collection.

The passages computed by IR-n are usually composed of a fixed number of sentences. Due to the format of the document collection, we have chosen a fixed number of words to limit the sentences, concretely, 30 words per sentence and 4 sentences per passage with overlapping of one passage. The IR-n system uses the Okapi similarity measure to calculate the weight of the words of the topic according to the document collection. More details about the IR-n system are given in [1]. The following section shows the logic form derivation task and the NLP rules that are applied to these logic forms with the aim of increasing some weights of the words of the topic. Finally, we describe each one of the submitted runs, our scores in the submitted runs and the conclusions.

^{*} This work has been partially supported by the Spanish Government (CICYT) with grant TIC2003-07158-C04-01.

2 Logic Form Derivation

In order to enhance the performance of our IR-n system, the logic form of the topics was computed. Each topic term in the logic form can modify its term weight according to its predicate type and the relationships between all the predicates in the logic form of the topic.

2.1 Logic Form Inference

The logic form of a sentence is calculated by means of a dependency relationship analysis between the words of the sentence. Our approach employs a set of NLP rules that infer several properties such as the predicate, the predicate type, the unique identifier of the predicate and the relationships between the different predicates in the logic form. This technique is different from other techniques such as Moldovan's [2] which constructs the logic form through the syntactic tree obtained from the output of the syntactic parser. Our logic form is based on the format of the logic form defined by the Logic Form Transformation of eXtended WordNet [4] lexical resource. As an example, the logic form "*story:NN(x14) of:IN(x14, x13) varian:NN(x10) NNC(x11, x10, x12) fry:NN(x12) and:CC(x13, x11, x6) emergency:NN(x5) NNC(x6, x5, x7) rescue:NN(x8) NNC(x7, x8, x9) committee:NN(x9) who:NN(x1) save:VB(e1, x1, x2) thousand:NN(x2) in:IN(e1, x3) marseille:NN(x3)*" is automatically inferred from the analysis of the dependency relationships between the words of the topic "*The story of Varian Fry and the Emergency Rescue Committee who saved thousands in Marseille*".

3 Applying Rules to Logic Forms to Increase the Topic Terms Weights

According to the processing of this logic form, when the predicate type corresponds to a preposition (*IN*) whose second argument is identified with an predicate whose type matches to noun (*NN*) or derives in a predicate whose type corresponds to noun, then the term weight associated to this last predicate is increased. This rule generally describes those grammatical utterances that have a circumstantial behaviour in the sentence (eg. in Marseille, in concentration camps, in Sweden, of Holocaust experience). So, we consider the nouns (*NN* predicate type) as very relevant words in the topic. As a result of this, the term weights of these words (terms) are increased by 15% of their initial values. Table 1 shows the term weights that the IR-n system associates to the topic "*The story of Varian Fry and the Emergency Rescue Committee who saved thousands in Marseille*". These terms are expressed by their stem form. The logic form inferred from this topic (see it in the previous section) has two predicates whose types are *IN*. The second argument of these predicates is instantiated to *x13* and *x3* respectively. *x13* derives in the predicates *x10*, *x12*, *x5*, *x8* and *x9* (details about derivation of predicates are provided in [3]) which types are *NN*, while the type of *x3* is directly *NN*. Then, according to this rule, these facts cause that the term weights associated with all these predicates increase their initial values.

Table 1. Terms weights assigned by IR-n system

Term (Porter-stem)	Initial Weight	Final Weight
stori	1.84449	1.84449
fry	6.19484	7.124066
emerg	6.47296	7.443904
rescu	6.19484	7.124066
committe	4.08194	4.694231
save	3.06725	3.06725
thousand	2.33944	2.33944
marseil	5.13363	5.9036745

4 Submitted Runs

This section describes the different submitted runs of our IR-n system. The differences between these five submitted runs are basically based on the treatment of the topics and on indexing a combination of different fields of the segments in the document collection. Thesaurus terms are used as keywords in the performance of the indexing and the searching processes. The following list shows the features of these five submitted runs according to the judgment pool priority order:

- **UA_TDN_FL_ASR06BA1A2 Run.** In this run IR-n system indexed the combination of the **ASRTEXT2006B**, **AUTOKEYWORD2004A1** and **AUTOKEYWORD2004A2** fields of the segments in the document collection. The English title, description and narrative topic fields were used in the construction of the queries. This was the only submitted run in which we applied the rules based on the processing of queries in the way of logic forms described in the previous section.
- **UA_TDN_ASR06BA1A2 Run.** In this run, as in UA_TDN_FL_ASR06BA1A2 run, IR-n system indexed the combination of the **ASRTEXT2006B**, **AUTOKEYWORD2004A1** and **AUTOKEYWORD2004A2** fields of the segments in the document collection. The English title, description and narrative topic fields were used in the construction of the queries.
- **UA_TD_ASR06BA2 Run.** In this run IR-n system indexed the combination of the **ASRTEXT2006B** and **AUTOKEYWORD2004A2** fields of the segments in the document collection. Only the title and description topic fields were used in the construction of the queries.
- **UA_TDN_ASR06BA2 Run.** In this run, as in previous run, IR-n system indexed the combination of the **ASRTEXT2006B** and **AUTOKEYWORD2004A2** fields of the segments in the document collection. The English title, description and narrative topic fields were used in the construction of the queries.
- **UA_TD_ASR06B Run.** In this required run, IR-n system only indexed the **ASRTEXT2006B** field of the segments in the document collection.

Table 2. Evaluation Results

run	map	R-prec	bpref	rr	p5	p20	p100	p1000
UA_TDN_FL_ASR06BA1A2	0.0369	0.0757	0.0651	0.1800	0.0882	0.1029	0.0821	0.0262
UA_TDN_ASR06BA1A2	0.0365	0.0714	0.0640	0.2255	0.0882	0.0956	0.0785	0.0260
UA_TD_ASR06BA2	0.0328	0.0727	0.0660	0.1681	0.1000	0.0868	0.0700	0.0264
UA_TDN_ASR06BA2	0.0345	0.0756	0.0691	0.1671	0.0765	0.0926	0.0774	0.0278
UA_TD_ASR06B	0.0339	0.0736	0.0792	0.2010	0.1235	0.1088	0.0709	0.0297

Only the title and description topic fields were used in the construction of the queries.

5 Results

Table 2 shows the results obtained by our system for each one of the submitted runs. These scores suggest that the use of the technique based on logic forms may improve the results of the statistical IR-n system.

6 Conclusions

In this year's CL-SR track at the CLEF 2006 workshop we have participated by applying our IR-n system to the English language. Our main aim is to evaluate the efficacy of the new treatment of topics by way of logic forms. According to the scores obtained we can conclude that when applying this processing technique (UA_TDN_FL_ASR06BA1A2) the results may be better than the ones obtained without the use of this new module (UA_TDN_ASR06BA1A2).

References

1. Llopis, F., Noguera, E.: Combining Passages in the Monolingual Task with the IR-n System. In: Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evaluation Forum, Vienna, Austria, pp. 204–207.
2. Moldovan, D.I., Rus, V.: Logic Form Transformation of WordNet and its Applicability to Question-Answering. In: Proceedings of 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France (2001)
3. Terol, R.M., Martínez-Barco, P., Palomar, M.: Applying Logic Forms to Biomedical Q-A. In: International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2005), Istanbul, Turkey, pp. 29–32 (2005)
4. Harabagiu, S., Miller, G.A., Moldovan, D.I.: WordNet 2 - A Morphologically and Semantically Enhanced Resource. In: Proceedings of ACL-SIGLEX99: Standardizing Lexical Resources, College Park, Maryland, pp. 1–8 (1999)

XML Information Retrieval from Spoken Word Archives

Robin Aly, Djoerd Hiemstra, Roeland Ordelman, Laurens van der Werff,
and Franciska de Jong

University of Twente, The Netherlands
{r.aly, d.hiemstra, r.j.f.ordelman, l.b.vanderwerff,
f.m.g.dejong}@utwente.nl

Abstract. In this paper the XML Information Retrieval System PF/Tijah is applied to retrieval tasks on large spoken document collections. The used example setting is the English CLEF-2006 CL-SR collection together with given English topics and self produced Dutch topics. The main findings presented in this paper are the easy way of adapting queries to use different kinds and combinations of metadata. Furthermore simple ways of combining different metadata kinds are shown to be beneficial in terms of mean average precision.

1 Introduction

Traditionally the retrieval of information from a multimedia document collection has been approached by creating indexes relying on manual annotation. With the emergence of digital archiving this approach is still widely in use and for many archiving institutes the creation of manually generated metadata is and will be an important part of the daily work. When the automation of metadata generation is considered, it is often seen as something that can enhance the existing process rather than replace it. The available metadata will therefore often be a combination of highly reliable and conceptually rich manual annotations, and (semi-) automatically generated metadata. A big challenge for Information Retrieval is to combine the various types of metadata and to let the user benefit from this combination.

Audio metadata may be regarded as a set of special document features; each feature adds to the overall representation of a document. Traditionally, these features include for example locating the speech/non-speech fragments, the identification of the speaker and production of full text transcripts of the speech within the document. In a lot of cases, speech recognition transcripts even provide the only means for searching a collection, for example when manual annotation of a collection is not feasible. More recently tools have become available for the extraction of additional speaker features from the audio signal such as emotions (e.g., affect bursts such as sobs, crying or words that express emotions), language, accent, dialect, age and gender.

Unlike with text content the notion of a document is not so obviously defined. A document, comprising for example a whole interview, can be very long

and probably not every part is relevant given a specific user request. Browsing manually through a long multimedia document to find the most relevant parts can be cumbersome. Therefore, multimedia retrieval systems have to be capable of retrieving only relevant *parts* of documents. As a consequence partitioning or segmentation of documents should take place prior to the actual extraction of features.

Usually, the segmentation algorithm for defining sensible document parts comes almost naturally with a document collection. In the English part of the Multilingual Access to Large Spoken Archives (MALACH) collection, which is the target collection for CLEF's CL-SR task, parts of interviews on a single coherent subject have been selected as segments. In a broadcast news retrieval system, single news items would be the obvious unit to choose for segmentation purposes.

However, given that the currently available techniques can generate multiple annotations, multiple segmentations should be supported. This complicates the task. For the MALACH collection an Information Retrieval system should be able to handle the following type of query: "give me document parts where male, native Dutch speakers are talking about [...] without expressing any emotions". Thus the retrieval system needs to consider various sources to yield a good performance. With this scenario the following questions need to be addressed: (i) how to efficiently store these multiple annotations, (ii) how to combine and rank retrieval scores based on them and (iii) how to select appropriate parts of the documents that are going to be returned as part of a result.

MPEG-7 [1] defines standardized ways of storing annotations of multimedia content. It has the potential of solving some of the issues related to the use of multiple annotations in multimedia retrieval. For this reason the underlying XML [2] format was chosen as a starting point for our research. An XML database is employed to store collections. For the information retrieval requests PF/Tijah, an XQuery extension made for this purpose, was used.

CLEF 2006 CL-SR is an ideal evaluation scenario for our research because: (i) The provided test collection existed in XML-like format, (ii) the collection was pre-segmented, (iii) each segment contained manual and automatic annotations, and (iv) these annotations could be combined in order to exploit their added value. CLEF 2006 CL-SR offers therefore a welcome framework for evaluating the preliminary implementation of our ideas described below. A second reason is that it links up to other work on spoken document retrieval for oral history collections done at the University of Twente, in particular to the CHoral project, which is focusing on speech retrieval techniques for Dutch cultural heritage collections. [3].

The rest of the paper is organized as follows. In section 2 the PF/Tijah module, its application to the retrieval task together with the cross-lingual application are presented. Section 3 presents our results on the given retrieval topics. We end the paper by giving a conclusion and future work in Section 4.

2 PF/Tijah: Text Search in XQuery

PF/Tijah is a research project run by the University of Twente with the goal to create an extendable search system. It is implemented as a module of the PathFinder XQuery (PF) execution system [4] and uses the Tijah XML Information Retrieval system [5]. PF/Tijah is part of the XML Database Management System MonetDB/XQuery. The whole system is an open source package developed in cooperation with CWI Amsterdam and the University of Munich. It is available from SourceForge.¹ The rest of this section is structured as follows: In Subsection 2.1 we describe the PF/Tijah system. The next Subsection 2.2 gives insights on application of the PF/Tijah system to perform the given tasks on the CLEF-2006 CL-SR collection. The last Subsection, 2.3, gives details on the cross-language aspects.

2.1 Features of PF/Tijah

PF/Tijah supports retrieving arbitrary parts of XML data, unlike traditional information retrieval systems for which the notion of a *document* needs to be defined up front by the application developer. For instance, if the data consists of MPEG-7 annotations, one can retrieve complete documents, but also only parts inside documents with no need to adapt the index or any other part of the system.

Information retrieval queries are specified in PF/Tijah through a user-defined function in XQuery that takes a NEXI query as its argument. NEXI [6] stands for *Narrowed Extended XPath*. It is narrowed since it only supports the descendant and the self axis steps (which selects the current context node - opposed to all nodes beneath it). “Extended” because of its special `about()` function that takes a sequence of nodes and ranks those by their estimated probability of relevance to the query consisting of a number of terms.

PF/Tijah supports extended result presentation by means of the underlying query language. For instance, when searching for document parts, it is easy to include any XML element in the collection in the result. This can be done in a declarative way (i.e., not focusing on how-to include something but on what to include). This could be additional information on the author of the document or technical information such as the encoding. The result is gathered using XPath expressions and is returned by means of XQuery element constructions.

As a result of the integration of information retrieval functionality in XQuery, PF/Tijah information retrieval search/ranking can be combined with traditional database query techniques, including for instance joins of datasets on certain values.

For example, one could search for video parts in which a word is mentioned, that is listed in a *name* attribute in an XML document collection on actors. It is also possible to narrow down the results by constraining the *actor* element further by, say, the earliest movie he played in.

¹ <http://sourceforge.net/projects/monetdb/>

2.2 Querying the CLEF-2006 CL-SR Collection

For a description of the anatomy of the CLEF-2006 CL-SR collection please refer to [7]. PF/Tijah can be used to query this collection as follows. Considering topic 1133, *“The story of Varian Fry and the Emergency Rescue Committee who saved thousands in Marseille”*, a simple CLEF CL-SR query that searches the ASR transcript version from 2004 is presented in Fig. 1. To understand the query it is necessary to have some basic knowledge of XQuery as for instance described by Katz et al. [8]. In the query, we specify which elements to search (in this case ASRTEXT2004A) explicitly. It is easy to run queries on other elements (annotations), for instance the SUMMARY elements, without the need to re-index. Furthermore, queries using several elements can be done by combining multiple about() functions with and or or operators in the NEXI query.

```
let $c := doc("Segments.xml")
for $doc in tijah-query($c, "//DOC[about(../ASRTEXT2004A,
    varian fry)];")
return $doc/DOCNO
```

Fig. 1. Simple PF/Tijah query

```
let $opt := <TijahOptions returnNumber="1000" algebraType="COARSE2"
    txtmodel_model="NLLR"/>
let $c := doc("Segments.xml")
(: for all topics :)
for $q in doc("EnglishQueries.xml")//top
    (: generate NEXI Query :)
    let $q_text := tijah-tokenize(exactly-one($q/title/text()))
    let $q_nexi := concat("//DOC[about(../ASRTEXT2004A,", $q_text, ")]");
    (: start query on the document node :)
    let $result := tijah-query-id($opt, $c, $q_nexi)
    (: for each result node print the standardized result :)
    for $doc at $rank in tijah-nodes($result)
    return
        string-join(($q/num/text(), "Q0", $doc/DOCNO/text(), string($rank),
            string(tijah-score($result, $doc)), "pftijah"), " ")
```

Fig. 2. PF/Tijah query that completely specifies a CLEF run

Interestingly, we can specify one complete CLEF CL-SR run in one huge XQuery statement as shown in Fig. 2. The query takes every topic from the CLEF topics file, creates a NEXI query from the topic title, runs each query on

the collection, and produces the official CLEF output format. The `<TijahOptions />` element contains options for processing the query, such as the retrieval model that is supposed to be used. More about PF/Tijah can be found in [9].

2.3 Cross-Lingual IR

For cross-language retrieval Dutch topics [7] were used as queries on the English collection. For this purpose all Dutch queries were fully automatically translated into English and then processed with the procedure described above.

As a machine translation (MT) system we used the free web translation service *freetranslation.com*², which is based on the SDL Enterprise Translation Server, a hybrid phrase based statistical and example based machine translation system. There was no pre- or post-processing carried out. The results using *freetranslation.com* seemed to be performing well, though no formal evaluation of the translation quality was carried out. Since online MT systems like this are being changed regularly, it will be difficult to repeat the achieved results. This is one of the reasons we are considering building our own statistical MT system for a future CLEF evaluation.

3 Experimental Results

We tested the out-of-the-box performance of PF/Tijah with provided, unmodified metadata. Table 1 shows the experimental results of 9 experiments. One experiment is the obligatory ASR run using relatively long queries consisting of of topic title and topic description (denoted as ‘TD’ in the table). The other eight runs were done using short queries with only the topic title (‘T’ in the table). Five out of nine runs use the original English topics (‘EN’ in the table). The four remaining runs use the manually translated Dutch topics (‘NL’ in the table) to search the English annotations. Runs were done on different annotations: on the ASR transcripts version 2004 (‘ASR04’), on the manual keywords (‘MK’) and/or on the manual summaries (‘SUM’).

Table 1 reports the LEF 2006 mean average precision results for 63 training topics and 33 test topics. The first five runs were officially submitted the other ones were self-assessed. The ASR-only results are significantly worse than any of the runs that use manual annotations (manual keywords or summaries of interview parts). Best results are obtained by combining the manual keywords and summaries. Interestingly, on the test topics, the difference between manual keywords only (UTmkEN) and manual keywords plus ASR (UTasr04mkEN) is statistically significant at the 5% level according to a paired sign test.

4 Conclusions and Future Work

Expectedly, retrieval results using ASR transcripts are far off from the results of manual annotation. For example, manually annotated summaries combined with

² <http://www.freetranslation.com/>

Table 1. Mean average precision (map) on train and test queries (slightly differs from official values due to fixed bugs)

Run name	Lang. Query Annotations			Train (map)	Test (map)
UTasr04aEN-TD	EN	TD	ASR04	0.069	0.047
UTasr04aEN	EN	T	ASR04	0.058	0.052
UTasr04aNL2	NL	T	ASR04	0.053	0.039
UTsummknENor	EN	T	MK, SUM	0.246	0.203
UTsummknNL2or	NL	T	MK, SUM	0.210	0.169
UTmkEN	EN	T	MK	0.201	0.151
UTmkNL2	NL	T	MK	0.163	0.119
UTasr04mkEN	EN	T	ASR04, MK	0.218	0.162
UTasr04mkNL2	NL	T	ASR04, MK	0.177	0.129

manual keywords have 0.2 mean average precision compared to 0.05 for the ASR transcripts. Taking costs and the available resources for the annotation process into account, manual annotation often turns out to be infeasible due to limited resources. Therefore, if no manual annotations are available – which is true for quite a number of audiovisual collections – automatic annotation at least allows a specific document to be considered if it contains relevant information to the user. In addition, experiments show that the combination of ASR and MK annotations improve the precision of the results significantly, compared to sole use of MK. Therefore we expect that the combination of manual and automatically generated annotations will be beneficial in general.

As it was stated in the introduction, using multiple annotations gives rise to several unsolved, important questions. The experiments described in this paper have been a first exploration of these issues using a well-defined spoken document retrieval evaluation corpus. We aim to continue this line of research by investigating (i) how annotation layers can most efficiently be stored, (ii) how retrieval scores could best be combined, and (iii) how document segmentation should be dealt with.

The storage structure used for the annotations was the predefined format (transformed to valid XML) of the CLEF-2006 CL-SR collection. All annotations were aligned after one segmentation scheme: parts of interviews on one coherent topic. This way all annotations nicely fitted into a rather flat hierarchical structure. In other also realistic settings segmentations could be more deeply nested or overlapping. For example, an automatically detected emotion could span multiple parts of an interview. This leaves the question open of how to efficiently store multiple annotations. Here, existing standards for data structures for the storage of annotations, like MPEG-7, have to be considered and evaluated.

The retrieval of information based on manual and automatic annotations was done so far using an “or” operator. This assumes that both annotation types contribute equally to the relevance of a segment. It is however more realistic to assume some kind of ordering or weighting. For example, manual keyword

annotations might receive a larger weight than findings in the ASR transcript. How to solve the problem of possible overlapping annotations during the combination of retrieval results on different annotations is still to be settled.

With possibly huge amounts of content in one single multimedia document a retrieval system has to be able to select only certain parts which are presented to the user as being relevant. Because annotations may segment a document in multiple ways, the issue of which part to return as relevant needs to be addressed in more detail.

Acknowledgements

We are grateful to the research programmes that made this work possible: Djord Hiemstra and Roeland Ordelman were funded by the Dutch BSIK project MultimediaN³. Robin Aly was funded by the Centre of Telematics and Information Technology's SRO-NICE programme⁴. Laurens van der Werff was funded by the Dutch NWO CATCH project Choral⁵. Many thanks to Arthur van Bunningen, Sander Evers, Robert de Groote, Ronald Poppe, Ingo Wassink, Dennis Reidsma, Dolf Trieschnigg, Lynn Packwood, Wim Fikkert, Jan Kuper, Dennis Hofs, Ivo Swartjes, Rieks op den Akker, Harold van Heerde, Boris van Schooten, Mariet Theune, Frits Ordelman, Charlotte Bijron, Sander Timmerman and Paul de Groot for translating the English topics to Dutch.

References

1. MPEG-Systems-Group: ISO/MPEG N4285, Text of ISO/IEC Final Draft International Standard 15938-1 Information Technology - Multimedia Content Description Interface - Part 1 Systems. ISO (2001)
2. Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E.: The extensible markup language (xml) 1.0 3rd edn. Technical report, W3C (2004)
3. Ordelman, R., de, F.J., van A.J.H.: The role of automated speech and audio analysis in semantic multimedia annotation. In: Proceedings of the International Conference on Visual Information Engineering (VIE2006), Bangalore, India (September 2006)
4. Boncz, P., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., Teubner, J.: MonetDB/XQuery: A fast XQuery processor powered by a relational engine. In: Proceedings of the ACM SIGMOD Conference on Management of Data, ACM, New York (2006)
5. List, J., Mihajlovic, V., Ramirez, G., de Vries, A., Hiemstra, D., Blok, H.: Tijah: Embracing IR methods in XML databases. *Information Retrieval Journal* 8(4), 547–570 (2005)
6. O'Keefe, R., Trotman, A.: The simplest query language that could possibly work. In: Proceedings of the 2nd Initiative for the Evaluation of XML Retrieval (INEX) (2004)

³ <http://www.multimedien.nl/>

⁴ <http://www.ctit.utwente.nl/research/sro/nice/>

⁵ <http://hmi.ewi.utwente.nl/project/CHoral>

7. Peters, C., Clough, P., Gey, F., Karlgen, J., Magnini, B., de Rijke, D.W.D.O., Stempfhuber, M. (eds.): Evaluation of Multilingual and Multi-modal Information Retrieval Seventh Workshop of the Cross-Language Evaluation Forum CLEF 2006, September 2006, Alicante, Spain. Springer, Heidelberg (to appear, 2007)
8. Katz, H., Chamberlin, D., Draper, D., Fernandez, M., Kay, M., Robie, J., Simeon, M.R.J., Tivy, J., Wadler, P.: XQuery from the Experts: A Guide to the W3c XML Query Language. Addison-Wesley, Reading (2003)
9. Hiemstra, D., Rode, H., van Os, R., Flokstra, J.: PFTijah: text search in an XML database system. In: Proceedings of 2nd International Workshop on Open Source Information Retrieval (OSIR) (2006)

Experiments for the Cross Language Speech Retrieval Task at CLEF 2006

Muath Alzghool and Diana Inkpen

School of Information Technology and Engineering
University of Ottawa
{alzghool,diana}@site.uottawa.ca

Abstract. This paper presents the second participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) task at CLEF 2006. We present the results of the submitted runs for the English collection and very briefly for the Czech collection, followed by many additional experiments. We have used two Information Retrieval systems in our experiments: SMART and Terrier, with several query expansion techniques (including a new method based on log-likelihood scores for collocations). Our experiments showed that query expansion methods do not help much for this collection. We tested different Automatic Speech Recognition transcripts and combinations. The retrieval results did not improve, probably because the speech recognition errors happened for the words that are important in retrieval. We present cross-language experiments, where the queries are automatically translated by combining the results of several online machine translation tools. Our experiments showed that high quality automatic translations (for French) led to results comparable with monolingual English, while the performance decreased for the other languages. Experiments on indexing the manual summaries and keywords gave the best retrieval results.

1 Introduction

This paper presents the second participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) track at CLEF 2006. We briefly describe the task [8]. Then, we present our systems, followed by results for the submitted runs for the English collection and very briefly for the Czech collection. We present results for many additional runs for the English collection. We experimented with many possible weighting schemes for indexing the documents and the queries, and with several query expansion techniques. We tested with different speech recognition transcripts to see if the word error rate has an impact on the retrieval performance. We describe cross-language experiments, where the queries are automatically translated from French, Spanish, German and Czech into English, by combining the results of several online machine translation (MT) tools. At the end we present the best results, when summaries and manual keywords were indexed.

2 System Description

The University of Ottawa Cross-Language Information Retrieval (IR) systems were built with off-the-shelf components. For translating the queries from French, Spanish,

German, and Czech into English, several free online machine translation tools were used. The idea behind using multiple translations is that they might provide more variety of words and phrases, therefore improving the retrieval performance. Seven online MT systems [4] were used for translating from Spanish, French, and German. We combined the outputs of the MT systems by simply concatenating all the translations. All seven translations of a title made the title of the translated query; the same was done for the description and narrative fields. We used the combined topics for all the cross-language experiments reported in this paper. For translation of the Czech language topics into English we were able to find only one online MT system.

For the retrieval part, the SMART [2,11] IR system and the Terrier [1,9] IR system were tested with many different weighting schemes for indexing the collection and the queries.

SMART was originally developed at Cornell University in the 1960s. SMART is based on the vector space model of information retrieval. we used mainly the *lnn.ntn* weighting scheme [2,11] which performs very well in CLEF-CLSR 2005 [4].

We have also used a query expansion mechanism with SMART, which follows the idea of extracting related words for each word in the topics using the Ngram Statistics Package (NSP) [10]. We extracted the top 6412 pairs of related words based on log likelihood ratios (high collocation scores in the corpus of ASR transcripts), using a window size of 10 words. We chose log-likelihood scores because they are known to work well even when the text corpus is small. For each word in the topics, we added the related words from this list of pairs. We call this approach SMART_{ns}.

Terrier was originally developed at University of Glasgow. It is based on Divergence from Randomness models (DFR) where IR is seen as a probabilistic process [1, 9]. We experimented with the *ln(exp)C2* weighting model, one of Terrier's DFR-based document weighting models.

We have also used a query expansion mechanism in Terrier, which follows the idea of measuring divergence from randomness. In our experiments, we applied the Kullback-Leibler (KL) model for query expansion [3, 9].

3 Experimental Results

3.1 Submitted Runs

Table 1 shows the results of the submitted results on the test data (33 queries). The evaluation measure we report is the standard measure computed with the *trec_eval* script (version 8): MAP (Mean Average Precision). The information about what fields of the topic were indexed is given in the column named Fields: T for title only, TD for title + description, TDN for title + description + narrative. For each run we include an additional description of the experimental settings and which document fields were indexed. For the *uoEnTDt04A06A* and *uoEnTDNtMan* runs we used the indexing scheme *ln(exp)C2* from Terrier; and for *uoEnTDNsQEx04*, *uoFrTDNs*, and *uoSpTDNs* we used the indexing scheme *lnn.ntn* from SMART. We used SMART_{ns} query expansion for the *uoEnTDNsQEx04* run, KL query expansion for

uoEnTDNtMan and uoEnTDt04A06A, and we didn't use any query expansion techniques for uoFrTDNs and uoSpTDNs.

Our required run, English TD (0.0565), obtained a lower result than our automatic English TDN run (0.0768), mainly due to different system settings, not due to the additional field N. Comparing the result of our required run to the best required run, submitted by Dublin City University (dcuEgTDauto, 0.0733) [6], their result was better with relative improvement 30%; but we obtained a comparable result using SMART with blind relevance feedback (SMARTnsp, 0.0754 – more details are given in section 3.2); our result was better with relative improvement 2%.

Table 1. Results of the five submitted runs, for topics in English, French, and Spanish. The required run (English, title + description) is in bold.

Runs for English	MAP	Fields	Description
uoEnTDNtMan	0.2902	TDN	Terrier: MANUALKEYWORD + SUMMARY
uoEnTDNsQEx04	0.0768	TDN	SMART: NSP query expansion ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoFrTDNs	0.0637	TDN	SMART: ASRTEXT2004A + AUTOKEY- WORD2004A1, A2
uoSpTDNs	0.0619	TDN	SMART: ASRTEXT2004A + AUTOKEY- WORD2004A1, A2
uoEnTDt04A06A	0.0565	TD	Terrier: ASRTEXT2004A + ASRTEXT2006A + AUTOKEYWORD2004A1, A2

We also participated in the task for Czech language. We indexed the Czech topics and ASR transcripts. Table 2 shows the results of the submitted runs on the test data (29 topics) for the Czech collection. The evaluation measure we report is the mean Generalized Average Precision (mGAP), which rewards retrieval of the right time-stamps in the collection. MAP scores could not be used because the speech transcripts were not segmented. We used the quickstart collection provided: each document contains 4-minute passages that start each minute (overlapping passages). From our results, we note:

- The mGAP is substantially low for all submitted runs.
- There is a small improvement when we indexed the field ENGLISH-MANUKEYWORD relative to the case when we indexed CZECHMANUKEYWORD.
- We got small improvements if CZECHMANUKEYWORD was added to the ASR field.
- Terrier’s results are slightly better than SMART’s for the required run.

Comparing our best run (0.0235) to the best run by University of West Bohemia team (0.0456) [5], our results were lower because we did not use any Czech-specific processing in our system (such as removing stop words or stemming), while in [5] this was done. In the rest of the paper we focus only on the English CLSR collection.

Table 2. Results of the five submitted runs for Czech collection. The required run (title + description) is in bold.

Runs for Czech	mGAP	Fields	Description
uoC-zEnTDNsMan	0.0235	TDN	SMART: ASRTEXT, CZECHAUTOKEYWORD, CZECHMANUKEYWORD, ENGLISH MANUKEYWORD, ENGLISHAUTOKEYWORD
uoCzTDNsMan	0.0200	TDN	SMART: ASRTEXT, CZECHAUTOKEYWORD, CZECHMANUKEYWORD
uoCzTDNs	0.0182	TDN	SMART: ASRTEXT, CZECHAUTOKEYWORD
uoCzTDs	0.0211	TD	SMART: ASRTEXT, CZECHAUTOKEYWORD
uoCzEnTDt	0.0218	TD	Terrier: ASRTEXT, CZECHAUTOKEYWORD

3.2 Comparison of Systems and Query Expansion Methods

Table 3 presents results for the best weighting schemes: for SMART we chose $\ln n$.ntn and for Terrier we chose the $\ln(\exp)C2$ weighting model, because they achieved the best results on the training data. We present results with and without relevance feedback. According to Table 3, we note that:

- Blind relevance feedback helps to improve the retrieval results in Terrier for TDN, TD, and T for the training data; the improvement was high for TD and T, but not for TDN. For the test data there is a small improvement.
- NSP relevance feedback with SMART does not help to improve the retrieval for the training data (except for TDN); the improvement on the test data was small.
- SMART results are better than Terrier results for the test data, but not for the training data.

Table 3. Results (MAP scores) for Terrier and SMART, with or without relevance feedback, for English topics. In bold are the best scores for TDN, TD, and T.

	System	Training			Test		
		TDN	TD	T	TDN	TD	T
1	SMART	0.0954	0.0906	0.0873	0.0766	0.0725	0.0759
	SMART _{nsf}	0.0923	0.0901	0.0870	0.0768	0.0754	0.0769
2	Terrier	0.0913	0.0834	0.0760	0.0651	0.0560	0.0656
	TerrierKL	0.0915	0.0952	0.0906	0.0654	0.0565	0.0685

3.3 Comparison of Retrieval Using Various ASR Transcripts

In order to find the best ASR transcripts to use for indexing the segments, we compared the retrieval results when using the ASR transcripts from the years 2003, 2004, and 2006 or combinations. We also wanted to find out if adding the automatic keywords helps to improve the retrieval results. The results of the experiments using Terrier and SMART are shown in Table 4 and Table 5, respectively. We note from the experimental results that:

Table 4. Results (MAP scores) for Terrier, with various ASR transcript combinations. In bold are the best scores for TDN, TD, and T.

Segment fields	Terrier					
	Training			Test		
	TDN	TD	T	TDN	TD	T
ASRTEXT 2003A	0.0733	0.0658	0.0684	0.0560	0.0473	0.0526
ASRTEXT 2004A	0.0794	0.0742	0.0722	0.0670	0.0569	0.0604
ASRTEXT 2006A	0.0799	0.0731	0.0741	0.0656	0.0575	0.0576
ASRTEXT 2006B	0.0840	0.0770	0.0776	0.0665	0.0576	0.0591
ASRTEXT 2003A+2004A	0.0759	0.0722	0.0705	0.0596	0.0472	0.0542
ASRTEXT 2004A+2006A	0.0811	0.0743	0.0730	0.0638	0.0492	0.0559
ASRTEXT 2004A+2006B	0.0804	0.0735	0.0732	0.0628	0.0494	0.0558
ASRTEXT 2003A+ AUTOKEYWORD2004A1,A2	0.0873	0.0859	0.0789	0.0657	0.0570	0.0671
ASRTEXT 2004A+ AUTOKEYWORD2004A1, A2	0.0915	0.0952	0.0906	0.0654	0.0565	0.0685
ASRTEXT 2006B+ AUTOKEYWORD2004A1,A2	0.0926	0.0932	0.0909	0.0717	0.0608	0.0661
ASRTEXT 2004A+2006A+ AUTOKEYWORD2004A1, A2	0.0915	0.0952	0.0925	0.0654	0.0565	0.0715
ASRTEXT 2004A+2006B+ AUTOKEYWORD2004A1,A2	0.0899	0.0909	0.0890	0.0640	0.0556	0.0692

Table 5. Results (MAP scores) for Terrier, with various ASR transcript combinations. In bold are the best scores for TDN, TD, and T.

Segment fields	SMART					
	Training			Test		
	TDN	TD	T	TDN	TD	T
ASRTEXT 2003A	0.0625	0.0586	0.0585	0.0508	0.0418	0.0457
ASRTEXT 2004A	0.0701	0.0657	0.0637	0.0614	0.0546	0.0540
ASRTEXT 2006A	0.0537	0.0594	0.0608	0.0455	0.0434	0.0491
ASRTEXT 2006B	0.0582	0.0635	0.0642	0.0484	0.0459	0.0505
ASRTEXT 2003A+2004A	0.0685	0.0646	0.0636	0.0533	0.0442	0.0503
ASRTEXT 2004A+2006A	0.0686	0.0699	0.0696	0.0543	0.0490	0.0555
ASRTEXT 2004A+2006B	0.0686	0.0713	0.0702	0.0542	0.0494	0.0553
ASRTEXT 2003A + AUTOKEYWORD2004A1,A2	0.0923	0.0847	0.0839	0.0674	0.0616	0.0690
ASRTEXT 2004A+ AUTOKEYWORD2004A1,A2	0.0954	0.0906	0.0873	0.0766	0.0725	0.0759
ASRTEXT 2006B+ AUTOKEYWORD2004A1,A2	0.0869	0.0892	0.0895	0.0650	0.0659	0.0734
ASRTEXT 2004A+ 2006A + AUTOKEYWORD2004A1,A2	0.0903	0.0932	0.0915	0.0654	0.0654	0.0777
ASRTEXT 2004A +2006B + AUTOKEYWORD2004A1,A2	0.0895	0.0931	0.0919	0.0652	0.0655	0.0742

- Using Terrier, the best field is ASRTEXT2006B which contains 7377 transcripts produced by the ASR system on 2006 and 727 transcripts produced by the ASR system in 2004, this improvement over using only the ASRTEXT2004A field is

very small. On the other hand, the best ASR field using SMART is ASRTEXT2004A.

- Any combination between two ASRTEXT fields does not help.
- Using Terrier and adding the automatic keywords to ASRTEXT2004A improved the retrieval for the training data but not for the test data. For SMART it helps for both the training and the test data.
- In general, adding the automatic keywords helps. Adding them to ASRTEXT2003A or ASRTEXT2006B improved the retrieval results for the training and test data.
- For the required submission run English TD, the maximum MAP score was obtained by the combination of ASRTEXT 2004A and 2006A plus autokeywords using Terrier (**0.0952**) or SMART (**0.0932**) on the training data; on the test data the combination of ASRTEXT 2004A and autokeywords using SMART obtained the highest value, **0.0725**, higher than the value we report in Table 1 for the submitted run.

3.4 Cross-Language Experiments

Table 6 presents results for the combined translation produced by the seven online MT tools, from French, Spanish, and German into English, for comparison with monolingual English experiments (the first line in the table). All the results in the table are from SMART using the lnn.ntn weighting scheme.

Since the result of combined translation for each language was better than when using individual translations from each MT tool on the CLEF 2005 CL-SR data [4], we used combined translations in our experiments.

The retrieval results for French translations were very close to the monolingual English results, especially on the training data. On the test data, the results were much worse when using only the titles of the topics, probably because the translations of the short titles were less precise. For translations from the other languages, the retrieval results deteriorate rapidly in comparison with the monolingual results. We believe that the quality of the French-English translations produced by online MT tools was very good, while the quality was lower for Spanish, German and Czech, successively.

Table 6. Results of the cross-language experiments, where the indexed fields are ASRTEXT2004A, and AUTOKEYWORD2004A1, A2 using SMART (lnn.ntn).

Language	Training			Test		
	TDN	TD	T	TDN	TD	T
English	0.0954	0.0906	0.0873	0.0766	0.0725	0.0759
French	0.0950	0.0904	0.0814	0.0637	0.0566	0.0483
Spanish	0.0773	0.0702	0.0656	0.0619	0.0589	0.0488
German	0.0653	0.0622	0.0611	0.0674	0.0605	0.0618
Czech	0.0585	0.0506	0.0421	0.0400	0.0309	0.0385

3.5 Manual Summaries and Keywords

Table 7 presents the results when only the manual keywords and the manual summaries were used. The retrieval performance improved a lot, for topics in all the languages. The MAP score jumped from 0.0654 to 0.2902 for English test data, TDN, with the $\ln(\exp)C2$ weighting model in Terrier. The results of cross-language experiments on the manual data show that the retrieval results for combined translation for French and Spanish language were very close to the monolingual English results on training data and test data. For all the experiments on manual summaries and keywords, Terrier's results are better than SMART's.

Our results on manual summaries and keywords for TD and TDN (0.2902, 0.2710) were better than the submitted runs by Dublin City University (0.2765, 0.2015) [6], the relative improvements for TD and TDN were 5% and 34% respectively.

Table 7. Results of indexing the manual keywords and summaries, using SMART with weighting scheme $\ln n.n\ln$, and Terrier with $\ln(\exp)C2$.

Language and System	Training			Test		
	TDN	TD	T	TDN	TD	T
English SMART	0.3097	0.2829	0.2564	0.2654	0.2344	0.2258
English Terrier	0.3242	0.3227	0.2944	0.2902	0.2710	0.2489
French SMART	0.2920	0.2731	0.2465	0.1861	0.1582	0.1495
French Terrier	0.3043	0.3066	0.2896	0.1977	0.1909	0.1651
Spanish SMART	0.2502	0.2324	0.2108	0.2204	0.1779	0.1513
Spanish Terrier	0.2899	0.2711	0.2834	0.2444	0.2165	0.1740
German SMART	0.2232	0.2182	0.1831	0.2059	0.1811	0.1868
German Terrier	0.2356	0.2317	0.2055	0.2294	0.2116	0.2179
Czech SMART	0.1766	0.1687	0.1416	0.1275	0.1014	0.1177
Czech Terrier	0.1822	0.1765	0.1480	0.1411	0.1092	0.1201

4 Conclusion

We experimented with two different systems: Terrier and SMART, with various weighting schemes for indexing the document and query terms. We proposed a new approach for query expansion that uses collocations with high log-likelihood ratio. Used with SMART, the method obtained a small improvement on test data (not statistically significant according to a Wilcoxon signed test). The KL blind relevance feedback method produced only small improvements with Terrier on test data. So, query expansion methods do not seem to help for this collection.

The improvements of mean word error rates in the ASR transcripts (of ASRTEXT2006A relative to ASRTEXT2004A) did not improve the retrieval results. Also, combining different ASR transcripts (with different error rates) did not help.

For some experiments, Terrier was better than SMART, for others it was not; therefore we cannot clearly choose one or another IR system for this collection.

The idea of using multiple translations proved to be good. More variety in the translations would be beneficial. The online MT systems that we used are rule-based

systems. Adding translations by statistical MT tools might help, since they could produce radically different translations.

On the manual data, the best MAP score we obtained is 0.2902, for the English test topics. On automatically-transcribed data the best result is 0.0766 MAP score. Since the improvement in the ASR word error rate does not improve the retrieval results, as shown from the experiments in section 3.3, we think that the justification for the difference to the manual summaries is due to the fact that summaries contain different words to represent the content of the segments. In future work we plan to investigate methods of removing or correcting some of the speech recognition errors in the ASR contents and to use speech lattices for indexing.

References

1. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems* 20(4), 357–389 (2002)
2. Buckley, C., Salton, G., Allan, J.: Automatic retrieval with locality information using SMART. In: *Text REtrieval Conference (TREC-1)*, pp. 59–72 (1993)
3. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)* 19(1), 1–27 (2001)
4. Inkpen, D., Alzghool, M., Islam, A.: Using various indexing schemes and multiple translations in the CL-SR task at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
5. Ircing, P., Müller, L.: The University of West Bohemia at CLEF 2006, the CL-SR track. In: *Evaluation of Multilingual and Multi-modal Information Retrieval Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September 19-21 (2006)*
6. Jones, G.J.F., Zhang, K., Lam-Adesina, A.M.: Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments. In: *Evaluation of Multilingual and Multi-modal Information Retrieval Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006, September 19-21, Alicante, Spain (2006)*
7. Oard, D.W., Soergel, D., Doermann, D., Huang, X., Murray, G.C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S.: Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In: *Proceedings of SIGIR (2004)*
8. Oard, D.W., Wang, J., J., Jones, G.J.F., White, R.W., Pecina, P., Soergel, D., Huang, X., Shafran, I.: Overview of the CLEF 2006 cross-language speech retrieval track. In: *Working Notes of the CLEF- 2006 Evaluation, Alicante, Spain*, p. 12 (2006)
9. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier Information Retrieval Platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) *ECIR 2005. LNCS*, vol. 3408, Springer, Heidelberg (2005), <http://ir.dcs.gla.ac.uk/wiki/Terrier>
10. Pedersen, T., Banerjee, S.: The design, implementation and use of the Ngram Statistics Package. In: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico (2003)*
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic retrieval. *Information Processing and Management* 24(5), 513–523 (1988)

CLEF-2006 CL-SR at Maryland: English and Czech

Jianqiang Wang¹ and Douglas W. Oard²

¹ Department of Library and Information Studies
State University of New York at Buffalo, Buffalo, NY 14260 USA
jw254@buffalo.edu

² College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
oard@umd.edu

Abstract. The University of Maryland participated in the English and Czech tasks. For English, one monolingual run using fields based on automatic transcription (the required condition) and one (otherwise identical) cross-language run using French queries were officially scored. Weighted use of alternative translations yielded an apparent improvement over one-best translation that was not statistically significant, but statistical translation models trained on European Parliament proceedings were found to be poorly matched to this task. Three contrastive runs in which manually generated metadata from the English collection was indexed were also officially scored. Results for Czech were not informative in this first year of that task.

1 Introduction

Previous experiments have shown that limitations in Automatic Speech Recognition (ASR) accuracy were an important factor degrading the retrieval effectiveness for spontaneous conversational speech [1,2]. In this year's CLEF CL-SR track, ASR text with a lower word error rate was provided for the same set of segmented English interviews used in last year's CL-SR track. Therefore, one of our goals was to determine the degree to which improved ASR accuracy could measurably improve retrieval effectiveness.

This year's track also introduced a new task, searching unsegmented Czech interviews. Unlike more traditional retrieval tasks, the objective in this case was to find points in the interviews that mark the beginning of relevant segments (i.e., points at which a searcher might wish to begin replay). A new evaluation metric, mean Generalized Average Precision (mGAP), was defined to evaluate system performance on this task. A description of mGAP and details on how it is computed can be found in the track overview paper. In this study, we were interested in evaluating retrieval based on overlapping passages, a retrieval techniques that has proven to be useful in document retrieval settings, while hopefully also gaining some insight into the suitability of the new evaluation metric.

2 Techniques

In this section, we briefly describe the techniques that we used.

2.1 Combining Evidence

Both test collections provide several types of data that are associated with the information to be retrieved, so we tried different pre-indexing combinations of that data. For the English test collection, we combined the ASR text generated with the 2004 system and the ASR text generated with the 2006 system, and we compared those results with results obtained by indexing each alone. Our expectation was that the two ASR systems would produce different errors, and that combining their results might therefore yield better retrieval effectiveness. Thesaurus terms generated automatically using a kNN classifier offer some measure of vocabulary expansion, so we added them to the ASR text combination as well.

The English test collection also contains a rich metadata produced by human indexers. Specifically, a set of (on average, five) thesaurus terms were manually assigned, and a three-sentence summary was written for each segment. In addition, the first names of persons mentioned in each segment are available, even when the name itself was not stated. We combined all three types of human-generated metadata to create an index for a contrastive condition.

The Czech document collection contains, in addition to ASR text, manually assigned English thesaurus terms, automatic translations of those terms into Czech, English thesaurus terms that were generated automatically using a kNN classifier that was trained on English segments, and automatic translations of those thesaurus terms into Czech. We tried two ways of combining this data. Czech translations of the automatically generated thesaurus terms were combined with Czech ASR text to produce the required automatic run. We also combined all available keywords (automatic and manual, English and Czech) with the ASR text, hoping that some proper names in English might match proper names in Czech.

2.2 Pruning Bidirectional Translations for Cross-Language Retrieval

In a previous study, we found that using several translation alternatives generated from bidirectional statistical translation could significantly outperform techniques that utilize only the most probable translation, but that using all possible translation alternatives was harmful because statistical techniques generate a large number of very unlikely translations [3]. The usual process for statistically deriving translation models is asymmetric, so we produce an initial bidirectional model by multiplying the translation probabilities from models trained separately in both directions and then renormalizing. This has the effect of emphasizing translation mappings that are well supported in both directions. Evidence for synonymy (in this case, from round trip translation using cascaded unidirectional mappings) is then used to aggregate probability mass

that would otherwise be somewhat eclectically divided across translation pairings that actually share similar meanings. To prune the resulting translations, document-language synonym sets that share the meaning of a query word are then arranged in decreasing order of renormalized translation probability, and a cumulative probability threshold is used to truncate that list. In our earlier study, we found that a cumulative probability threshold of 0.9 typically results in near-peak cross-language retrieval effectiveness. Therefore, in this study, we tried two conditions: one-best translation (i.e., a threshold of zero) and multiple alternatives with a threshold of 0.9.

We used the same bilingual corpus that we used in last year's experiment. Specifically, to produce a statistical translation table from French to English, we used the freely available GIZA++ toolkit [4]¹ to train translation models with the Europarl parallel corpus [5]. Europarl contains 677,913 automatically aligned sentence pairs in English and French from the European Parliament. We stripped accents from every character and filtered out implausible sentence alignments by eliminating sentence pairs that had a token ratio either smaller than 0.2 or larger than 5; that resulted in 672,247 sentence pairs that were actually used. We started with 10 IBM Model 1 iterations, followed by 5 Hidden Markov Model (HMM) iterations, and ending with 5 IBM Model 4 iterations. The result is a three-column table that specifies, for each French-English word pair, the normalized translation probability of the English word given the French word. Starting from the same aligned sentence pairs, we used the same process to produce a second translation table from French to English. Due to the time and resource constraints, we were not able to try a similar technique for Czech this year. All of our processing for Czech was, therefore, monolingual.

3 English Experiment Results

The one required run for the CLEF-2006 CL-SR track called for use of the *title and description fields* as a basis for formulating queries. To maximize comparability across our runs we therefore used all words from those fields as the query (a condition we call "TD") for our five official submissions. Stopwords in each query (and in each segment) were automatically removed (after translation) by InQuery, which was the retrieval system used for all of our experiments. Stemming of the queries (after translation) and segments was performed automatically by InQuery using kstem. Statistical significance is reported for $p < 0.05$ by a Wilcoxon signed rank test for paired samples.

3.1 Officially Scored Runs

Table 1 shows the experiment conditions and the Mean Average Precision (MAP) for the five official runs that we submitted. Not surprisingly, every run based on manually created metadata statistically significantly outperformed every run based on automatically generated fields.

¹ <http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html>

Table 1. MAP for official runs, TD queries

run name	umd.auto	umd.auto.fr0.9	umd.manu	umd.manu.fr0.9	umd.manu.fr0
	monolingual	CL-SR	monolingual	CL-SR	CL-SR
doc fields	ASR 2004A	ASR 2004A	NAME	NAME	NAME
	ASR 2006A	ASR 2006A	manualkey	manualkey	manualkey
	autokey04A1A2	autokey04A1A2	SUMMARY	SUMMARY	SUMMARY
MAP	0.0543	0.0209	0.235	0.1026	0.0956

As Table 1 shows, our one officially scored automatic CLIR run, umd.auto.fr0.9 (with a threshold of 0.9), achieved only 38% of the MAP of the corresponding monolingual MAP, a statistically significant difference. We suspect that domain mismatch between the corpus used for training statistical translation model and the document collection might be a contributing factor, but further investigation of this hypothesis is needed. A locally scored one-best contrastive condition (not shown) yielded about the same results.

Not surprisingly, combining first names, segment summaries, and manually assigned thesaurus terms produced the best retrieval effectiveness (as measured by MAP). Among the three runs with manually created metadata, umd.manu.fr0.9 and umd.manu.fr0 are a pair of comparative cross-language runs: umd.manu.fr0.9 had a threshold of 0.9 (which usually led to the selection of multiple translations), while umd.manu.fr0 used only the most probable French translation for each English query word. These settings yielded 44% and 41% of the corresponding monolingual MAP, respectively. The 7% apparent relative increase in MAP with a threshold of 0.9 (compared with one-best translation) was not statistically significant.

3.2 Additional Locally Scored Runs

In addition to our officially scored monolingual run umd.auto that used four automatically generated fields (ASRTEXT2004A, ASRTEXT2006A, AUTOKEYWORD2004A1, and AUTOKEYWORD2004A2), we scored three runs based on automatically generated data locally: umd.asr04a, umd.asr06a, and umd.asr06b. These runs used only the ASRTEXT2004A, ASRTEXT2006A, or ASRTEXT2006B fields, respectively (ASRTEXT2006A is empty for some segments; in ASRTEXT2004B those segments are filled in with words from ASRTEXT2004A). Table 2 shows the MAP for each of these runs and (again) for umd.auto. Combining the four automatically generated fields yielded a slight apparent improvement in MAP over any run that used only one of those four fields,

Table 2. Mean average precision (MAP) for 4 monolingual automatic runs, TD queries

run name	umd.auto	umd.asr04a	umd.asr06a	umd.asr06b
MAP	0.0543	0.0514	0.0517	0.0514

although the differences are not statistically significant. Interestingly, there is no noticeable difference among umd.asr04a, umd.asr06a and umd.asr06b even when average precision is compared on a topic-by-topic basis, despite the fact that the ASR text produced by the 2006 system is reported to have a markedly lower word error rate than that produced by the 2004 system.

3.3 Detailed Failure Analysis

To investigate the factors that could have had a major influence on the effectiveness of runs with automatic data, we conducted a topic-by-topic comparison of average precision. To facilitate that analysis, we produced two additional runs with queries that were formed using words from the title field only, one using ASRTEXT2006B only (AUTO) and the other the three manually constructed metadata fields (MANU). We focused our analysis on those topics that have a MAP of 0.2 or larger from the MANU run for consistency with the failure analysis framework that we have applied previously. With this constraint applied, 15 topics remain for analysis. Figure 1 shows the topic-by-topic comparison of average precision. As we saw in 2006, the difference in average precision between the two conditions is quite large for most topics.

Using title-only queries allowed us to look at the contribution of each query word in more detail. Specifically, we looked at the number of segments in which

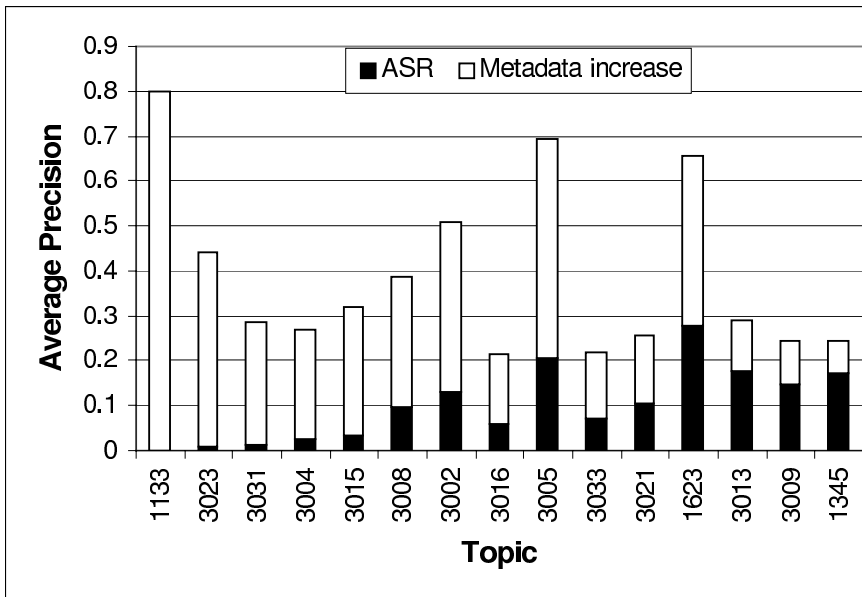


Fig. 1. Query-by-query comparison of retrieval effectiveness (average precision) between ASR text and metadata, 15 title queries with average precision of metadata equal to or higher than 0.2. in increasing order of (ASR MAP)/(metadata MAP)

Table 3. Query word statistics in the document collection. ASR#: the number of ASR segments that the query word appears; Metadata#: the number of segments in which the query word appears in manual metadata. All statistics are after stemming

Topic	Word	ASR#	Metadata#	Topic	Word	ASR#	Metadata#
1133	varian	0	4	3005	death	287	1013
	fry	4	4		marches	1327	736
3023	attitudes	60	242	3033	immigration	153	325
	germans	4235	1254		palestine	188	90
3031	activities	305	447	3021	survivors	724	169
	dp	0	192		contributions	50	6
	camps	3062	2452		developing	121	26
3004	deportation	277	568		israel	641	221
	assembly	59	9	1623	jewish	4083	1780
	points	1144	16		partisans	469	315
3015	mass	107	302		poland	1586	3694
	shootings	580	82	3013	yeshiva	101	23
3008	liberation	594	609		poland	1586	3694
	experience	811	658	3009	jewish	4083	1780
3002	survivors	724	169		children	2551	426
	impact	35	107		schools	2812	448
	children	2551	426	1345	bombing	593	71
	grandchildren	219	90		birkenau	167	677
3016	forced	654	1366		buchenwald	139	144
	labor	399	828				
	making	2383	80				
	bricks	197	12				

query word appears (i.e., the “document frequency”). As Table 3 shows, there are some marked differences in the prevalence of query term occurrences between automatically and manually generated fields. In last year’s study, poor relative retrieval effectiveness with automatically-generated data was most often associated with a failure to recognize at least one important query word (often a person or location name). This year, by contrast, we see only two obvious cases that fit that pattern (“varian” and “dp”). This may be because proper names are less common in this year’s topic titles. Instead, the pattern that we see is that query words actually appear quite often in ASR segments, and in some cases perhaps too often. Said another way, our problem last year was recall; this year our problem seems to be precision. We’ll next need to actually read some of the ASR text to see how well this supposition is supported. But it is intriguing to note that the failure pattern this year seems to be very different from last year’s.

4 Czech Experiment Results

We were also one of three teams to participate in the first year of the Czech evaluation. Limited time prevented us from implementing any language processing

Table 4. mGAP for the three official Czech runs, TD queries

run name	umd.all	umd.asr	umd.akey.asr
mGAP	0.0070	0.0066	0.0060

specific to Czech, so all of our runs were based on string matching without any morphological normalization. Automatic segmentation into overlapping passages was provided by the organizers, and we used that segmentation unchanged. Our sole research question for Czech was, therefore, whether the new start-time evaluation metric yielded results that could be used to compare variant systems.

We submitted three officially scored runs for Czech: umd.all used all available fields (both automatically and manually generated, in both English and Czech), umd.asr used only the Czech ASR text, and umd.akey.asr used both ASR text and the automatic Czech translations of the automatically generated thesaurus terms. Table 4 shows the resulting mGAP values. About all that we can conclude from these results is that they do not serve as a basis for making meaningful comparisons. We have identified three possible causes that merit investigation: (1) our retrieval system may indeed be performing poorly, (2) the evaluation metric may have unanticipated weaknesses, (3) the relevance assessments may contain errors. There is some reason to suspect that the problem may lie with our system design, which contains two known weaknesses. Most obviously, Czech is a highly inflected language in which some degree of morphological normalization is more important than it would be, for example, for English. Good morphological analysis tools are available for Czech, so this problem should be easily overcome. The second known problem is more subtle: overlapping segmentation can yield redundant highly ranked segments, but the mGAP scoring process penalized redundancy. Failing to prune redundant segments prior to submission likely resulted in a systematic reduction in our scores. It is not clear whether these two explanations together suffice to explain the very low reported scores this year, but the test collection that is now available from this year’s Czech task is exactly the resource that we need to answer that question.

5 Conclusion

Earlier experiments with searching broadcast news yielded excellent results, leading to a plausible conclusion that searching speech was a solved problem. As with all such claims, that is both true and false. Techniques for searching broadcast news are now largely well understood, but searching spontaneous conversational speech based solely on automatically generated transcripts remains a very challenging task. We continue to be surprised by some aspects of our English results, and we are only beginning to understand what happens when we look beyond manually segmented English to unsegmented Czech.

References

1. Oard, D.W., Soergel, D., Doermann, D., Huang, X., Murray, G.C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S., Mayfield, J., Kharevych, L., Strassel, S.: Building an information retrieval test collection for spontaneous conversational speech. In: *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 38–41. ACM Press, New York (2004)
2. Wang, J., Oard, D.W.: Clef-2005 cl-sr at maryland: Document and query expansion using side collections and thesauri. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
3. Wang, J., Oard, D.W.: Combining bidirectional translation and synonymy for cross-language information retrieval. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 202–229. ACM Press, New York (2006)
4. Och, F.J., Ney, H.: Improved statistical alignment models. In: *ACL'00*, Hongkong, China, pp. 440–447 (2000)
5. Koehn, P.: Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft (2002)

Dublin City University at CLEF 2006: Cross-Language Speech Retrieval (CL-SR) Experiments

Gareth J.F. Jones, Ke Zhang, and Adenike M. Lam-Adesina

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
{gjones,kzhang,adenike}@computing.dcu.ie

Abstract. The Dublin City University participation in the CLEF 2006 CL-SR task concentrated on exploring the combination of the multiple fields associated with the documents. This was based on use of the extended BM25F field combination model originally developed for multi-field text documents. Additionally, we again conducted runs with our existing information retrieval methods based on the Okapi model. This latter method required an approach to determining approximate sentence boundaries within the free-flowing automatic transcription provided, to enable us to use our summary-based pseudo relevance feedback (PRF). Experiments were conducted for the English document collection with topics translated into English using Systran V3.0 machine translation.

1 Introduction

The Dublin City University participation in the CLEF 2006 CL-SR task concentrated on exploring the combination of the multiple fields associated with the speech documents. It is not immediately clear how best to combine the diverse fields of this document set most effectively in ad hoc information retrieval (IR), such as the CLEF 2006 CL-SR task. Our study is based on using the document field combination extended version of BM25 termed BM25F introduced in [1]. In addition, we carried out runs using our existing IR methods based on the Okapi model with summary-based pseudo-relevance feedback (PRF) [2]. Our official submissions included both English monolingual and French bilingual tasks using automatic only and combined automatic and manual fields. Topics were translated into English using the Systran V3.0 machine translation system. The resulting translated English topics were applied to the English document collection.

The remainder of this paper is structured as follows: Section 2 summarises the motivation and implementation of the BM25F retrieval model, Section 3 overviews our basic retrieval system and describes our sentence boundary creation technique, Section 4 presents the results of our experimental investigations, and Section 5 concludes the paper with a discussion of our results.

2 Field Combination

The “documents” of the speech collection are based on sections of extended interviews which are segmented into topically related sections. The spoken documents are provided with a rich set of data fields, full details of these are given in [3]. In summary the fields comprise:

- a transcription of the spoken content of the document generated using an automatic speech recognition (ASR) system,
- two assigned sets of keywords generated automatically (AKW1,AKW2),
- one assigned set of manually generated keywords (MKW1),
- a short three sentence manually written summary of each segment,
- a list of the names of all individuals appearing in the segment.

Two standard methods of combining multiple document fields in retrieval are:

- to simply merge all the fields into a single document representation and apply standard single document field information retrieval methods,
- to index the fields separately, perform individual retrieval runs for each field and then to merge the resulting ranked lists by summing in a process of data fusion.

The topic of field combination for this type of task with ranked information retrieval schemes is explored in [1]. That paper demonstrated the weaknesses of the simple standard combination methods and proposed an extended version of the standard BM25 term weighting scheme referred to as BM25F, which combines multiple fields in a more well founded way.

The BM25F combination approach uses a simple weighted summation of the multiple fields of the documents to form a single field for each document in the usual way. The importance of each document field for retrieval can be determined empirically in separate runs for each field, the count of each term appearing in each field is multiplied by a scalar constant representing the importance of this field, and the components of all fields are then summed to form the overall single field document representation for indexing. Once the fields have been combined in a weighted sum, standard single field IR methods can be applied.

3 System Setup

The basis of our experimental system is the City University research distribution version of the Okapi system [4]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffixes are stripped using the Okapi implementation of Porter stemming [5] and terms are indexed using a small standard set of synonyms. None of these procedures were adapted for the CLEF 2006 CL-SR test collection.

Our experiments augmented the standard Okapi retrieval system with two variations of PRF based on extensions of the Robertson selection value (rsv) for expansion term selection. One method is a novel field-based PRF which we

are currently developing [6], and the other a summary-based method [7] used extensively in our earlier CLEF submissions.

3.1 Term Weighting

Document terms were weighted using the Okapi BM25 weighting scheme developed in [4]. With collection frequency weight $cfw(i) = \log(((rload + 0.5)(N - n(i) - bigrload + rload + 0.5)) / ((n(i) - rload + 0.5)(bigrload - rload + 0.5)))$, and $rload = 4$ and $bigrload = 5$. The BM25 k_1 and b values used for our submitted runs were tuned using the CLEF 2005 training and test topics.

3.2 Pseudo-relevance Feedback

The main challenge for query expansion is the selection of appropriate terms from the assumed relevant documents. For the CL-SR task our query expansion method operates as follows.

Field-Based PRF. Query expansion based on the standard Okapi relevance feedback model makes no use of the field structure of multi-field documents. We are currently exploring possible methods of making use of field structure to improve the quality of expansion term selection. For this current investigation we adopted the following method.

The fields are merged as described in the previous section and documents retrieved using the initial query. The Robertson selection value (rsv) [4] is then calculated separately for each field of the original document, but where the document position in the ranked retrieval list has been determined using the combined document. The rsv values in the ranked lists for each field are then normalised with respect to the highest scoring term in each list, and then summed to form a single merged rsv list from the expansion terms are selected. The objective of this process is to favour the selection of expansion terms which are ranked highly by multiple fields, rather than those which may obtain a high rsv value based on their association with a minority of the fields.

Summary-Based PRF. The method used here is based on our work originally described in [7], and modified for the CLEF 2005 CL-SR task [2].¹ A summary is made of the ASR transcription of each of the top ranked documents, which are assumed to be relevant for each PRF. Each document summary is then expanded to include all terms in the other metadata fields used in this document index. All non-stopwords in these augmented summaries are then ranked using a slightly modified version of the rsv [4]. In our modified version of $rsv(i)$, potential expansion terms are selected from the augmented summaries of the top ranked documents, but ranked using statistics from a larger number of assumed relevant ranked documents from the initial run.

¹ We refer to this as “Summary-Based PRF” for historical reasons, the “summaries” used here are unrelated to those provided with the CL-SR test collection.

Sentence Selection. The summary-based PRF method operates by selecting topic expansion terms from document summaries. However, since the ASR transcriptions of the conversational speech documents do not contain punctuation, we developed a method of selecting significant document segments to identify documents “summaries”. This uses a method derived from Luhn’s word cluster hypothesis. Luhn’s hypothesis states that significant words separated by not more than 5 non-significant words are likely to be strongly related. Clusters of these strongly related significant word were identified in the running document transcription by searching for word groups separated by not more than 5 insignificant words. Words appearing between clusters are not included in clusters, but can be ignored for the purposes of query expansion since they are by definition stop words. The clusters were then awarded a significance score based on two measures:

- Luhn’s Keyword Cluster Method: Luhn’s method assigns a significance score to each word cluster.
- Query-Bias Method: Assigns a score to each word cluster based on the number of query terms in each one.

The overall score for each word cluster was then formed by summing these two measures for each one. The word clusters were then ranked by score with the highest scoring ones selected as the segment summary.

4 Experimental Investigation

This section gives results of our experimental investigations for the CLEF 2006 CL-SR task. We first present results for our field combination experiments and then those for experiments using our summary-based PRF method. Results are shown for precision at rank cutoff 5, 10 and 30 documents, standard mean average precision (MAP) and the total number of relevant documents retrieved summed across all topics down to rank of 1000 documents.

For our formal submitted runs the system parameters were selected by optimising results for the CLEF 2005 CL-SR training and test collection. Our submitted runs for the CLEF 2006 are indicated by a * in the tables.

4.1 Field Combination Experiments

Two sets of experiments were carried out using the field combination method. The first uses all the document fields combining manual and automatically generated fields, and the other only the automatically generated fields. We report results for our formal submitted runs using our field-based PRF method and also baseline results without feedback. We also give further results obtained by optimising performance for our systems using the CLEF 2006 CL-SR test set.

Table 1. All fields results with parameters set using CLEF 2005 data

	TD	Recall	MAP	P5	P10	P30
Monolingual English	Baseline:	1844	0.223	0.366	0.293	0.255
	PRF*	1864	0.202	0.321	0.288	0.252
Bilingual French-English	Baseline	1491	0.158	0.306	0.256	0.204
	PRF*	1567	0.160	0.291	0.252	0.199

Table 2. All fields results with parameters optimised using CLEF 2006 topics

	TD	Recall	MAP	P5	P10	P30
Monolingual English	Baseline:	1908	0.234	0.364	0.342	0.303
	PRF	1929	0.243	0.364	0.370	0.305
Bilingual French-English	Baseline	1560	0.172	0.315	0.267	0.231
	PRF	1601	0.173	0.315	0.267	0.225

Table 3. Auto only fields results with parameters set using CLEF 2005 data

	TD	Recall	MAP	P5	P10	P30
Monolingual English	Baseline	1290	0.071	0.163	0.163	0.149
	PRF*	1361	0.073	0.152	0.142	0.146
Bilingual French-English	Baseline	1070	0.047	0.119	0.113	0.106
	PRF*	1097	0.047	0.106	0.094	0.102

Table 4. Auto only fields results with parameters optimised using CLEF 2006 topics

	TD	Recall	MAP	P5	P10	P30
Monolingual English	Baseline	1335	0.080	0.224	0.215	0.169
	PRF	1379	0.094	0.188	0.206	0.184
Bilingual French-English	Baseline	1110	0.050	0.121	0.127	0.123
	PRF	1167	0.055	0.127	0.142	0.124

All Field Experiments

Submitted Runs. Based on development runs with the CLEF 2005 data the Okapi parameters were set empirically as follows: $k_1 = 6.2$ and $b = 0.4$, and document fields were weighted as follows: Name field $\times 1$; Manualkeyword field $\times 10$; Summary field $\times 10$; ASR2006B $\times 2$; Autokeyword1 $\times 1$; Autokeyword2 $\times 1$.

The contents of each field were multiplied by the appropriate factor and summed to form the single field document for indexing. Note the unusually high value of k_1 arises due to the change in the $tf(i, j)$ profile resulting from the summation of the document fields [1].

Results for English monolingual and bilingual French-English runs are shown in Table 1. For both topic sets the top 20 ranked terms were added to the topic for the PRF run with the original topic terms upweighted by 3.0.

It can be seen from these results that, as is usually the case for cross-language IR, performance for monolingual English is better than bilingual French-English for all measures. A little more surprising is that while the application of the field-based PRF gives a small improvement in the number of relevant documents retrieved, there is little effect on MAP, and precision at high ranked cut off points is generally degraded. PRF methods for this task are the subject of ongoing research, and we will be exploring these results further.

Further Runs. Subsequent to the release of the relevance set for the CLEF 2006 topic set further experiments were conducted to explore the potential for improvement in retrieval performance when the system parameters are optimised for the topic set. We next show our best results achieved so far using the field combination method. For these runs the fields were weighted as follows: Name field $\times 1$; Manualkeyword field $\times 5$; Summary field $\times 5$; ASR2006B $\times 1$; Autokeyword1 $\times 1$; Autokeyword2 $\times 1$, and the Okapi parameters set empirically as follows: $k_1 = 10.5$ and $b = 0.35$.

The results for English monolingual and bilingual French-English runs are shown in Table 2. For monolingual English the top 20 terms were added to the topic for PRF run with the original topic terms upweighted by 33.0 For the French bilingual runs the top 60 terms were added to the topic with the original terms upweighted by 20.0. For these additional runs all test topics were included in all cases.

Looking at these additional results it can be seen that parameter optimisation gives a good improvement in all measures. It is not immediately clear whether this arises due to the instability of the parameters of our system, or a difference in some feature of the topics between CLEF 2005 and CLEF 2006. We will be investigating this issue further. Performance between monolingual English and bilingual French-English is similar to that observed for the submitted runs. PRF is generally more effective or neutral with the revised parameters, again we plan to conduct further exploration of PRF for multi-field documents to better understand these results.

Automatic Only Field Experiments

Submitted Runs. Based on development runs with the CLEF 2005 CL-SR data the system parameters for the submitted automatic only field experiments were set empirically as follows: $k_1 = 5.2$ and $b = 0.2$ and the document fields were weighted as follows: ASR2006B $\times 2$; Autokeyword1 $\times 1$; Autokeyword2 $\times 1$.

These were again summed to form the single field document for indexing. The same French-English topic translations were used for the automatic only field experiments as for the all field experiments.

The results of English monolingual and bilingual French-English runs are shown in Table 3. The top 30 terms were added to the topic for PRF run with the original topic terms upweighted by 3.0. From these results it can again be seen that there is the expected reduction in performance between monolingual English and bilingual French-English. The field-based PRF is once again shown

Table 5. Results for monolingual English with all document fields

		Recall	MAP	P10	P30
TDN	Baseline	1871	0.279	0.449	0.367
	PRF ^T	1867	0.291	0.461	0.377
TD	Baseline	1819	0.241	0.436	0.321
	PRF	1885	0.259	0.415	0.348

Table 6. Results for monolingual English with auto document fields

		Recall	MAP	P10	P30
TDN	Baseline	1259	0.070	0.182	0.160
	PRF	1252	0.072	0.194	0.155
TD	Baseline	1247	0.058	0.124	0.122
	PRF	1255	0.066	0.146	0.129

generally not to be effective with this dataset. There is a small improvement in the number of relevant documents retrieved, but no there is little positive impact for precision.

Further Runs. Subsequent to the release of the relevance set for the CLEF 2006 topic set further experiments were again conducted with system system parameters optimised using the test data set. For these runs the field weights were identical to those above for the submitted runs, and the Okapi parameters were modified as follows: $k_1 = 40.0$ and $b = 0.3$.

The results of English monolingual and bilingual French-English runs are shown in Table 4. For monolingual English the top 40 terms were added to the topic for PRF run with the original topic terms upweighted by 3.0 For the bilingual French-English runs the top 60 terms were added to the topic with the original terms upweighted by 3.5.

Once again optimising the system parameters results in an improvement effectiveness of the PRF method. Further experiments are planned to explore these results further.

4.2 Summary-Based PRF Experiments

For these experiments the document fields were combined into a single field for indexing without application of the field weighting method. The Okapi parameters were again selected using the CLEF 2005 CL-SR training and test collections. The values were set as follows: $k_1=1.4$ and $b=0.6$. For all our PRF runs, 3 documents were assumed relevant for term selection and document summaries comprised the best scoring 6 word clusters. The *rsv* values to rank the potential expansion terms were estimated based on the top 20 ranked assumed relevant documents. The top 40 ranked expansion terms taken from the clusters were added to the original query in each case. Based on results from our previous

experiments in CLEF, the original topic terms are up-weighted by a factor of 3.5 relative to terms introduced by PRF.

All Field Experiments. Table 5 shows results for monolingual English retrieval based on all document fields using TDN and TD field topics.² Rather surprisingly the baseline result for this system is better than either of those using the field-weighted method shown in Table 1. The summary-based PRF is shown to be effective here, as it has in many previous submissions to CLEF tracks in previous years. TDN topics produce better results than the TD only topics.

Automatic Only Field Experiments. Table 6 shows results for monolingual English retrieval based on only auto document fields again using TDN and TD topic fields. The summary-based PRF method is again effective for this document index. However, the results are not as good as those in Table 3 using the field weighted document combination method.

5 Conclusions and Further Work

This paper has described results for our participation in the CLEF 2006 CL-SR track. We explored use of a field combination method for multi-field documents, and two methods of PRF. Results indicate that further exploration is required of the field combination approach and our new field-based PRF method. Our existing summary-based PRF method is shown to be effective for this task. Further work will also concentrate on combining effective elements of both methods.

References

- [1] Robertson, S.E., Zaragoza, H., Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields. In: Proceedings of the 13th ACM International Conference on Information and Knowledge Management, pp. 42–49 (2004)
- [2] Lam-Adesina, A.M., Jones, G.J.F.: Dublin City University at CLEF 2005: Cross-Language Speech Retrieval (CL-SR) Experiments. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [3] White, R.W., Oard, D.W., Jones, G.J.F., Soergel, D., Huang, X.: Overview of the CLEF-2005 Cross-Language Speech Retrieval Track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [4] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference (TREC-3), NIST pp. 109–126 (1995)

² Our submitted run using this method was found to be in error, and the corrected one is denoted by the [†] in this table.

- [5] Porter, M.F.: An Algorithm for Suffix Stripping. Program 14, 10–137 (1980)
- [6] Zhang, K.: Cross-Language Spoken Document Retrieval from Oral History Archives, MSc Dissertation, School of Computing, Dublin City University (2006)
- [7] Lam-Adesina, A.M., Jones, G.J.F.: Applying Summarization Techniques for Term Selection in Relevance Feedback. In: Proceedings of the Twenty-Fourth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, pp. 1–9. ACM Press, New York (2001)

Overview of WebCLEF 2006

Krisztian Balog¹, Leif Azzopardi³, Jaap Kamps^{1,2}, and Maarten de Rijke¹

¹ ISLA, University of Amsterdam

² Archive and Information Studies, University of Amsterdam

`kbalog,kamps,mdr@science.uva.nl`

³ Department of Computing Science, University of Glasgow

`leif@dcs.gla.ac.uk`

Abstract. We report on the CLEF 2006 WebCLEF track devoted to crosslingual web retrieval. We provide details about the retrieval tasks, the used topic set, and the results of the participants. WebCLEF 2006 used a stream of known-item topics consisting of: (i) manual topics (including a selection of WebCLEF 2005 topics, and a set of new topics) and (ii) automatically generated topics (generated using two techniques). The results over all topics show that current CLIR systems are very effective, retrieving on average the target page in the top ranks. Manually constructed topics result in higher performance than and automatically generated ones. And finally, the resulting scores on automatic topics provide a reasonable ranking of the systems, showing that automatically generated topics are an attractive alternative in situations where manual topics are not readily available.

1 Introduction

The web presents one of the greatest challenges for cross-language information retrieval [5]. Content on the web is essentially multilingual, and web users are often polyglots. The European web space is a case in point: the majority of Europeans speak at least one language other than their mother-tongue, and the Internet is a frequent reason to use a foreign language [4]. The challenge of crosslingual web retrieval is addressed, head-on, by WebCLEF [12].

The crosslingual web retrieval track uses an extensive collection of spidered web sites of European governments, baptized EUROGOV [8]. The retrieval task at WebCLEF 2006 is based on a stream of known-item topics in a range of languages. This task, which is labeled *mixed-monolingual retrieval*, was pioneered at WebCLEF 2005 [9]. Participants of WebCLEF 2005 expressed the wish to be able to iron out issues with the systems they built during that year's campaign, since for many it was their first attempt at web IR with lots of languages, encoding issues, different formats, and noisy data. The continuation of this known-item retrieval task at WebCLEF 2006 allows veteran participants to take stock and make meaningful comparisons of their results over years. To facilitate this, we decided to include a selection of WebCLEF 2005 topics in the topic set (also

available for training purposes), as well as a set of new known-item topics. Also, we decided to trial the automatic generation of known-item topics [2]. By contrasting manually developed topics with automatically generated topics, we hope to gain insight in the validity of the automatically generated topics, especially in a multilingual environment. Our main findings are the following. First, the results over all topics show that current CLIR systems are quite effective, retrieving on average the target page in the top few ranks. Second, the manually constructed topics result in higher performance than the automatically generated ones. Third, the resulting scores on automatic topics give, at least, a solid indication of performance, and can hence be an attractive alternative in situations where manual topics are not readily available.

The remainder of this paper is structured as follows. Section 2 gives the details of the method for automatically generating known-item topics. Next, in Section 3, we discuss the details of the track set-up: the retrieval task, document collection, and topics of request. Section 4 reports the runs submitted by participants, and Section 5 discusses the results of the official submissions. Finally, in Section 6 we discuss our findings and draw some initial conclusions.

2 Automatic Topic Construction

This year we experimented with the automatic generation of known-item topics. The main advantage of automatically generating queries is that for any given test collection numerous queries of varying styles and quality can be produced at minimal cost [2]. In the WebCLEF setting this could be especially rewarding, since manual development of topics on all the different languages would require specialized human resources. The aim of this trial is to determine whether such topics are comparable to the manual topics with respect to the ranking of systems based on these topics. The following subsection describes how known item topics can be automatically generated using a generative process, along with the details of the topics generated for WebCLEF.

2.1 Known Item Topic Generation

To create simulated known item topics, we model the following behavior of a known-item searcher. We assume that the user wants to retrieve a particular document that they have seen before in the collection, because some need has arisen calling for this document. The user then tries to re-construct or recall terms, phrases and features that would help identify this document, which they pose as a query. The basic algorithm that we use for generating queries was introduced by (**author?**) [2], and is based on an abstraction of the actual querying process. It was described as follows:

- Initialize an empty query $q = \{\}$
- Select the document d to be the known-item with probability $p(d)$
- Select the query length k with probability $p(k)$

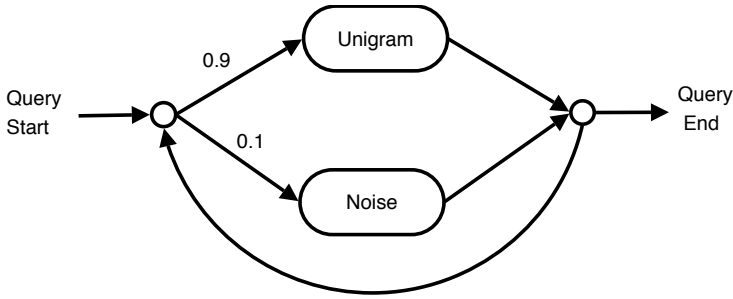


Fig. 1. The process of *auto-uni* query generation

- Repeat k times:
 - Select a term t from the document model of d with probability $p(t|\theta_d)$
 - Add t to the query q .
- Record d and q to define the known-item/query pair.

By repeatedly performing this algorithm we can create many queries. Before doing so, the probability distributions $p(d)$, $p(k)$ and $p(t|\theta_d)$ need to be defined. By using different probability distributions we can characterize different types and styles of queries that a user may submit.

(author?) [2] conducted experiments on an English test collection using various term sampling methods in order to simulate different styles of queries. In one case, they set the probability of selecting a term from the document model to a uniform distribution, where $p(t|\theta_d)$ was set to zero for all terms that did not occur in the document, whilst all other terms were assigned an equal probability. Compared to other types of queries, they found that using a uniform selection produced queries which were the most similar to real queries in terms of the performance and ranking of three different retrieval models.

In the construction of a set of known-item topics for the EUROGOV collection, we also use uniform sampling. However, we have tried to incorporate some realism into the querying process by including query noise, and then phrase extraction. Query noise can be thought of as the terms that users submit which do not appear in the known item. This may be because of poor memory or incorrect terms being recalled. To include some noise to the process of generating the queries, our model for sampling query terms is broken into two parts: sampling from the document (in our case uniformly) and sampling terms at random (i.e., noise). Figure 1 shows the sampling process; where a term is drawn from the unigram document model with some probability λ , or it is drawn from the noise model with probability $1 - \lambda$. This is similar to the query generation process described in [7] for Language Models. Consequently, as λ tends to one, we assume that the user has almost perfect recollection of the original document. Conversely, as λ tends to zero, we assume that the user's memory of the document

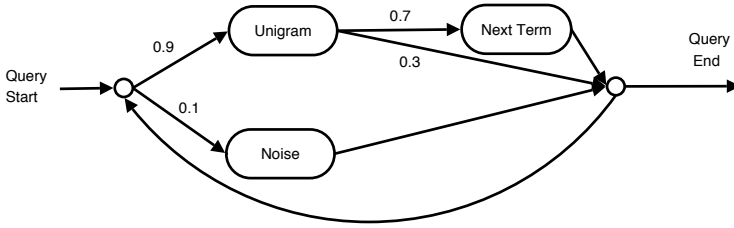


Fig. 2. The process of *auto-bi* query generation

degrades to the point that they know the document exists but they have no idea as to the terms other than randomly selecting terms (from the collection). We used $\lambda = 0.9$ for topic generation as analysis of the queries found that approximate 10% of query terms were noisy. The probability of a term given the noisy distribution was set to the probability of the term occurring in the collection. This model was used for our first setting, called *auto-uni*.

We further extended the process of sampling terms from a document. Once a term has been sampled from the document, we assume that there is some probability that the subsequent term in the document will be drawn. For instance given the sentence, "... Information Retrieval Agent ...," if the first term sampled is "Retrieval", then the subsequent term selected will be "Agent." This was included to provide some notion of phrase extraction to the process of selecting query terms. The process is depicted in Figure 2. This model was used for our second setting, called *auto-bi*, where we either add the subsequent term with $p = 0.7$, or not with probability $(1 - p) = 0.3$.

We indexed each domain within the EuroGOV collection separately, using the Lemur language modeling toolkit [6]. We experimented with two different styles of queries, and for each of them we generated 30 queries per top level domain. For both settings, the query length k was selected using a Poisson distribution where the mean was set to three, which reflected the average query length of manual queries. However, two restrictions were placed on sampled query terms: (i) the size of a term must contain more than three characters, and (ii) the term must not contain any numeric characters. Finally, the document prior $p(d)$ was set to a uniform distribution.

To wrap up, we summarize the process used. A document was randomly selected from the collection, and query terms were drawn either from the document unigram itself indiscriminately, or the noise unigram. These were the *auto-uni* topics, and for the *auto-bi* topics an additional step was introduced to include bi-gram selections. While these two models are quite simple, the results from these initial experiments are promising, and motivate further work with more sophisticated models for topic generation. A natural step would be to take structure and document priors into account. (For instance, do users really want to retrieve documents with equal probability? Do users draw query terms from certain parts of the document? Etc)

3 The WebCLEF 2006 Tasks

3.1 Document Collection

For the purposes of the WebCLEF track the EUROGOV corpus was developed [8], a crawl of European government-related sites, where collection building is less restricted by intellectual property rights. It is a multilingual web corpus, which contains over 3.5 million pages from 27 primary domains, covering over twenty languages. There is no single language that dominates the corpus, and its linguistic diversity provides a natural setting for multilingual web search.

3.2 Topics

The topic set for WebCLEF 2006 consists of a stream of 1,940 known-item topics, made up of both manual and automatically generated topics. As is shown in Table 1, a total of 195 manual topics were re-used from WebCLEF 2005, and 125 new manual topics were constructed. For the generated topics, we focused on 27 primary domains and generated 30 topics using the auto-uni query generation, and another 30 topics using the auto-bi query generation (see Section 2 for details), amounting to 810 automatic topics for each of the methods.

After the runs had been evaluated, we observed that the performance achieved on the automatic topics was frequently quite poor. We found that in several cases none of the participants found any relevant page within the top 50 returned results. These are often mixed-language topics, a result of language diversity within a primary domain, or they proved to be too hard for another reason.

In our post-submission analysis we decided to zoom in on a subset of the topics by removing any topics that did not meet the following criterion: whether any participant found the targeted page within the top 50.

Table 1 presents the number of original, deleted and remaining topics. 820 out of the 1,940 original topics were removed. Most of the removed topics are automatic (803), but there are also a few manual ones (17). The remaining topic set contains 1,120 topics, and is referred as the *new topic set*.

We decided to re-evaluate the submitted runs using this *new* topic set. Since it is a subset of the original topic collection, participants did not have to make any efforts. Submitted runs were re-evaluated using a restricted version of the (original) qrels that correspond to the new topic set.

Table 1. Number of topics in the *original topic set*, and in the *new topic set* where we only retain topics for which at least one participant retrieved the relevant page

	all	auto	auto-uni	auto-bi	manual	manual-o	manual-n
original	1,940	1,620	810	810	320	195	125
new	1,120	817	415	402	303	183	120
deleted	820	803	395	408	17	12	5

3.3 Retrieval Task

WebCLEF 2006 saw the continuation of the *Mixed Monolingual* task from WebCLEF 2005 [9]. This task is meant to simulate a user searching for a known-item page in a European language. The mixed-monolingual task uses the title field of the topics to create a set of monolingual known-item topics.

Our emphasis this year is on the mixed monolingual task. The manual topics in the topic set contain an English translation of the query. Hence, using only the manual topics, experiments with a *Multilingual* task are possible. This task is meant to simulate a user looking for a certain known-item page in a particular European language. The user, however, uses English to formulate her query. This multilingual task used the English translations of the original topic statements.

3.4 Submission

For each task, participating teams were allowed to submit up to 5 runs. The results had to be submitted in TREC format. For each topic a ranked list of no more than 50 results should be returned. For each topic at least 1 result must be returned. Participants were also asked to provide a list of the metadata fields they used, and a brief description of the methods and techniques employed.

3.5 Evaluation

The WebCLEF 2006 topics were known-item topics where a unique URL is targetted (unless there are page-duplicates in the collection, or near duplicates). Hence, we opted for a precision based measure. The main metric used for evaluation was *mean reciprocal rank* (MRR). The reciprocal rank is calculated as one divided by the rank at which the (first and in this case only) relevant page is found. The mean reciprocal rank is obtained by averaging the reciprocal ranks of a set of topics.

4 Submitted Runs

There were 8 participating teams that managed to submit official runs to WebCLEF 2006: BUAP (buap), University of Indonesia (depok), the University of Hildesheim (hildesheim), the Open Text Corporation (hummingbird), the University of Amsterdam (isla), the University of Salamanca (reina), the Universidad Politècnica de Valencia (rfia), and the Universidad Complutense Madrid (ucm). For details of the respective retrieval approaches to crosslingual web retrieval, we refer to the participants' papers.

Table 2 lists the runs submitted to WebCLEF 2006: 35 for the mixed-monolingual task, and 1 for the bilingual task. We also indicate the use of topic metadata, either the topic's language (TL), the targetted page's language (PL), or the targetted page's domain (PD). The mean reciprocal rank (MRR) is reported

Table 2. Summary of all runs submitted to WebCLEF 2006. The ‘metadata usage’ columns indicate usage of topic metadata: topic language (TL), page language (PL), page domain (PD). Mean Reciprocal Rank (MRR) scores are reported for the original and new topic set. For each team, its best scoring non-metadata run is in italics, and its best scoring metadata run is in boldface. Scores reported at the Multilingual section are based only on the manual topics.

Group id	Run name	Metadata usage		topics	
		TL	PL	PD	original
<i>Monolingual runs:</i>					
buap	<i>allpt40bi</i>			Y	<i>0.0157 0.0272</i>
depok	UI1DTA	Y			0.0404 0.0699
	UI2DTF	Y			0.0918 0.1589
	UI3DTAF	Y			0.0253 0.0439
	UI4DTW	Y			0.0116 0.0202
hildesheim	UHi1-5-10			Y	0.0718 0.1243
	UHi510			Y	0.0718 0.1243
	UHiBase			Y	0.0795 0.1376
	UHiBrf1			Y	0.0677 0.1173
	UHiBrf2			Y	0.0676 0.1171
	UHiTitle			Y	0.0724 0.1254
hummingbird	humWC06				0.1133 0.1962
	<i>humWC06dp</i>				<i>0.1209 0.2092</i>
	humWC06dpc				0.1169 0.2023
	humWC06dpcD			Y	0.1380 0.2390
	humWC06p				0.1180 0.2044
isla	<i>Baseline</i>				<i>0.1694 0.2933</i>
	Comb				0.1685 0.2918
	CombMeta			Y	0.1947 0.3370
	CombNboost			Y	0.1954 0.3384
	CombPhrase			Y	0.2001 0.3464
reina	usal_base	Y			0.0100 0.0174
	usal_mix	Y			0.0137 0.0237
	USAL_mix_hp	Y	Y		0.0139 0.0241
	usal_mix_hp	Y			0.0139 0.0241
	usal_mix_hp-ok	Y			0.0139 0.0241
rfia	DPSinDiac	Y		Y	0.0982 0.1700
	ERConDiac	Y		Y	0.1006 0.1742
	ERFinal	Y		Y	0.1021 0.1768
	ERSinDiac	Y		Y	0.1021 0.1768
ucm	webclef-run-all-2006	Y			0.0870 0.1505
	webclef-run-all-2006-def-ok	Y			0.0870 0.1505
	webclef-run-all-2006-def-ok-2	Y			0.0870 0.1505
	webclef-run-all-2006-ok-conref	Y			0.0870 0.1505
	webclef-run-all-OK-definitivo	Y			0.0870 0.1505
<i>Multilingual runs:</i>					
hildesheim	<i>UHiMu</i>				<i>0.2553 0.2686</i>

Table 3. Breakdown of the number of topics over domains for each topic type (*new topic set*)

	AT	BE	CY	CZ	DE	DK	EE	ES	EU	FI	FR	GR	HU	IE
auto_uni	20	4	20	11	4	26	29	27	17	3	10	24	5	17
auto_bi	23	7	20	6	5	23	19	20	15	9	14	28	7	17
auto	43	11	40	17	9	49	48	47	32	12	24	52	12	34
manual_old	1	2	1		21	9		27	27	1	1		14	1
manual_new	3	1			28			23	4	1			11	1
manual	4	3	1		49	9		50	31	2	1		25	2
all	47	14	41	17	58	58	48	97	63	14	25	52	37	36

	IT	LT	LU	LV	MT	NL	PL	PT	RU	SE	SI	SK	UK
auto_uni	17	21	17	12	20	10	20	5	5	7	23	21	20
auto_bi	21	17	18	10	21	10	18	6	5	14	22	13	14
auto	38	38	35	22	41	20	38	11	10	21	45	34	34
manual_old				2	2	29	1	25	8				11
manual_new						28	1						19
manual				2	2	57	2	25	8				30
all	38	38	35	24	43	77	40	36	18	21	45	34	64

over both the original and the new topic set. The official results of WebCLEF 2006 were based on the original topic set containing 1,940 topics. As detailed in Section 3.2 above, we have pruned the topic set by removing topics for which none of the participants retrieved the target page, resulting in 1,120 topics. In Appendix A, we provide scores for various breakdowns for both the original topic set and the new topic set.

In Table 3, a breakdown of the 1,120 topics in the new topic set is given. The topics cover 27 primary domains, and the number of topics per domain varies between 14 and 97.

The task description stated that for each topic, at least 1 result must be returned. Several runs did not meet this condition. The best results per team were achieved using 1 or more metadata fields. Knowledge of the page’s primary domain (shown in the PD column in Table 2) seemed moderately effective.

5 Results

This year our focus is on the Mixed-Monolingual task. A large number of topics were made available, consisting of old manual, new manual, and automatically generated topics. Evaluation results showed that the performance achieved on the automatic topics are frequently very poor, and we made a new topic set where we removed topics for which none of the participants found any relevant page within the top 50 returned results. All the results presented in this section correspond to the new topic set consisting of 1,120 topics.

Table 4. Best overall results using the *new topic set*. The results are reported on *all* topics, the *automatic* and *manual* subsets of topics, and *average* is calculated from the auto and manual scores.

Group id	Run	all	auto	manual	average
isla	combPhrase	0.3464	0.3145	0.4411	0.3778
hummingbird	humWC06dpcD	0.2390	0.1396	0.5068	0.3232
depok	UI2DTF	0.1589	0.0923	0.3386	0.2154
rfia	ERFinal	0.1768	0.1556	0.2431	0.1993
hildesheim	UHiBase /5-10	0.1376	0.0685	0.3299	0.1992
ucm	webclef-run-all-2006-def-ok-2	0.1505	0.1103	0.2591	0.1847
buap	allpt40bi	0.0272	0.0080	0.0790	0.0435
reina	USAL_mix_hp	0.0241	0.0075	0.0689	0.0382

5.1 Mixed-Monolingual

We look at each team’s best scoring run, independent of whether it was a baseline run or used some of the topic metadata. Table 4 presents the scores of the participating teams. We report the results over the whole new qrel set (*all*), and over the *automatic* and *manual* subsets of topics. The automatic topics proved to be more difficult than manual ones. This may be due in part to the fact that the manual topics cover 11 languages, but the generated topics cover all 27 domains in EUROGOV including the more difficult domains and languages. Another important factor may be the imperfections in the generated topics. Apart from the lower scores, the auto topics also dominate the manual topics in number. Therefore we also used the average of the auto and manual scores for ranking participants. Defining an overall ranking of teams is not straightforward, since one team may outperform another on the automatic topics, but perform worse on the manual ones. Still, we observe that participants can be unambiguously assigned into one out of three bins based on either the *all* or the *average* scores: the first bin consisting of hummingbird and isla; the second bin of depok, hildesheim, rfia, and ucm; and the third bin of buap and reina.

Figure 3 shows the relative performance of systems over all topics, over the manual topics, and over the automatic topics. Since there are some notable differences in score, and zoom in on the scores over automatic and manual topics.

5.2 Evaluation on Automatic Topics

Automatic topics were generated using two different methods, as described in Section 2 above. The participating teams’ scores did not show significant variance between the difficulty of topics, using the two generators. Table 5 provides details of the best runs when evaluation is restricted to automatically generated topics only.

Note that the scores included in Table 5 are measured on the new topic set. Notice, by the way, that there is very little difference between the number of topics within the *new topic set* for the two automatic topic subsets (*auto-uni* and *auto-bi* in Table 1).

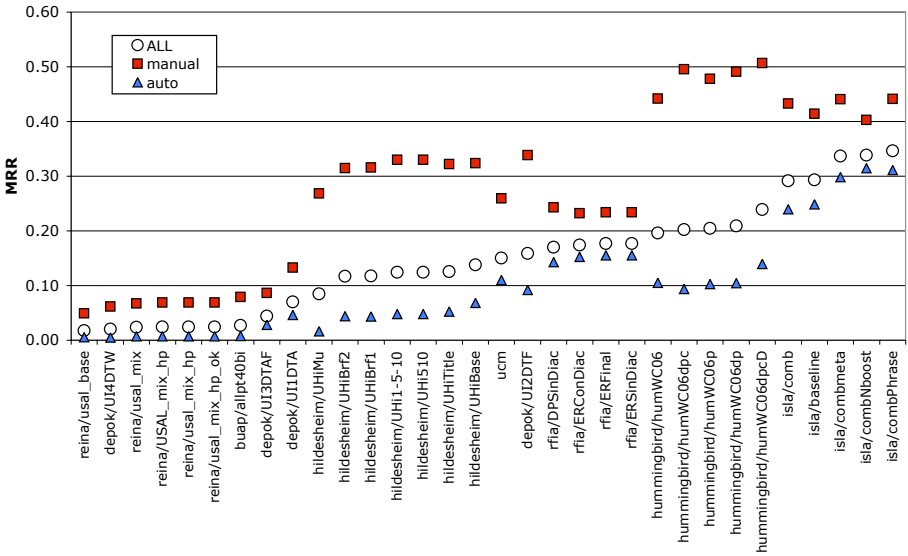


Fig. 3. Performance of all submitted runs on auto, manual, and all topics (*new topic set*)

Table 5. Best runs using the automatic topics in the *new topic set*

Group id	Run	auto	auto-uni	auto-bi
isla	combNboost	0.3145	0.3114	0.3176
rfa	ERFinal	0.1556	0.1568	0.1544
hummingbird	humWC06dpcD	0.1396	0.1408	0.1384
ucm	webclef-run-all-2006	0.1103	0.1128	0.1077
depok	UI2DTF	0.0923	0.1024	0.0819
hildesheim	UHiBase	0.0685	0.0640	0.0731
buap	allpt40bi	0.0080	0.0061	0.0099
reina	USAL_mix_hp	0.0075	0.0126	0.0022

In general, the two query generation methods perform very similarly, and it is system specific whether one type of automatic topics is preferred over the other. Our initial results with automatically generated queries are promising, but still a large portion of these topics are not realistic. This motivates us to work further on more advanced query generation methods.

5.3 Evaluation on Manual Topics

The manual topics include 183 old and 120 new queries. Old topics were randomly sampled from last year’s topics, while new topics were developed by *Universidad Complutense de Madrid (UCM)* and the track organizers. The new

Table 6. Best manual runs using the *new topic set*

Group id	Run	manual	old	new
hummingbird	humWC06dpcD	0.5068	0.4936	0.5269
isla	combPhrase	0.4411	0.3822	0.5310
depok	UI2DTF	0.3386	0.2783	0.4307
hildesheim	UHi1-5-10	0.3299	0.2717	0.4187
ucm	webclef-run-all-2006-def-ok-2	0.2591	0.2133	0.3289
rfia	DPSinDiac	0.2431	0.1926	0.3201
buap	allpt40bi	0.0790	0.0863	0.0679
reina	USAL_mix_hp	0.0689	0.0822	0.0488

Table 7. Kendall tau rank correlation, two-sided *p*-value

		all	auto	auto-uni	auto-bi	manual	manual-new	manual-old
all	τ	0.8182	0.7726	0.8125	0.5935	0.6292		0.5707
	p	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000
auto	τ		0.9412	0.9688	0.4108	0.4575		0.3945
	p		0.0000	0.0000	0.0006	0.0001		0.0010
auto-uni	τ			0.9097	0.3717	0.4183		0.3619
	p			0.0000	0.0019	0.0005		0.0025
auto-bi	τ				0.4029	0.4762		0.3800
	p				0.0008	0.0000		0.0016
manual	τ					0.9123		0.9642
	p					0.0000		0.0000
manual-new	τ							0.8769
	p							0.0000

topics cover only languages for which expertise was available: Dutch, English, German, Hungarian, and Spanish.

In case of the old manual topics we witnessed improvements for all teams that took part in WebCLEF 2005, compared to their last year's scores. Moreover, we found that most participating systems performed better on the new manual topics, compared to the old ones. A possible explanation is the nature of the topics, namely the new topics may be more appropriate for know-item search. Also, language coverage of the new manual topics could play a role.

5.4 Comparing Rankings

To compare the rankings of systems using the different topic sets we used Kendall's τ correlation, which has been previously used for comparing rankings of systems [10, 11]. The systems defined by their runs, are ordered by MRR for each topic set (i.e. the manual topics vs. automatic topics), and the two rankings are compared. If there is a high correlation between the manual and automatic topics then this would provide strong evidence to suggest that automatic queries can be used to predict the rankings of systems (w.r.t. manual topics).

We found that a weak to moderate, but statistically significant, positive correlation between automatic and manual topics existed ($\tau \approx 0.4$). Table 7 reports the Kendall τ correlation given each topic set against the other.

The rankings resulting from the topics generated with the “auto-bi” method are somewhat more correlated with the manual rankings than the ranking resulting from the topics generated with the “auto-uni” method. On the other hand, a very strong positive correlation ($\tau \approx 0.8 - 1.0$) is found between the ranking of runs obtained using new manual topics and the ranking of runs resulting from using old manual topics. Note that the new topic set we introduced did not affect the relative ranking of systems, thus the correlation scores we reported here are exactly the same for the *original* and for the *new topic sets*.

To provide some context for the correlations of system rankings, for ad-hoc retrieval, (author?) [10] used pseudo relevance assessments, in lieu of relevance assessments, as a way to simulate the assessments. The correlation between the ranking with the actual assessments and the pseudo assessments was around $\tau \approx 0.4 - 0.5$ on the TREC3-8 collections. In contrast, (author?) [11] found that using relevance assessments created by two different human assessors for the same set of topics, had a τ correlation of 0.938 on TREC4.

The gap between pseudo/generated and manual in terms of the correlation in ranking systems appears to be quite large. However, in our case the generation of automatic topics can be refined in order to model more accurately the process of topic generation. It is anticipated that this would lead to an improved correlation with manual topics. The fact that the correlation of generated topics is even moderate given the simplicity of the models is very encouraging.

5.5 Performance per Language

Table 8 scores the average score over all systems, broken down over the 27 domains in the topic set. We see considerable variation in average score over all topics, ranging from 0.0196 (Czech Republic) to 0.2883 (Netherlands). The average scores for the automatic topics range from 0.196 (Czech Republic) to 0.1452 (Italy). The average scores for the manual topics range from 0.0153 (Cyprus) to 0.6191 (Belgium).

5.6 Multilingual Runs

Our main focus this year was on the monolingual task, but we allowed submissions for multilingual experiments within the mixed-monolingual setup. The manual topics (both old and new ones) are provided with English titles. The automatically generated topics do not have English translations.

We received only one multilingual submission, from the *University of Hildesheim*. The evaluation of the multilingual run is restricted to the manual topics in the topic set, Table 2 summarizes the results of that run. A detailed breakdown over the different topic types is provided in Appendix A (Tables 9 and 10)

Table 8. Average MRR of the submitted runs, by domain and topic type (*new topic set*)

	AT	BE	CY	CZ	DE	DK	EE	ES	EU	FI
auto_uni	0.0693	0.1294	0.1062	0.0129	0.0653	0.1143	0.1006	0.1694	0.0713	0.0905
auto_bi	0.1264	0.0530	0.1166	0.0318	0.0317	0.1426	0.0861	0.0456	0.0321	0.1317
auto	0.0998	0.0808	0.1114	0.0196	0.0466	0.1276	0.0948	0.1167	0.0529	0.1214
manual_old	0.2917	0.5988	0.0153		0.1610	0.3187		0.2967	0.1835	0.0277
manual_new	0.4325	0.6597			0.3938			0.1540	0.2966	0.2777
manual	0.3973	0.6191	0.0153		0.2940	0.3187		0.2311	0.1981	0.1527
all	0.1252	0.1961	0.1090	0.0196	0.2556	0.1572	0.0948	0.1757	0.1244	0.1259

	FR	GR	HU	IE	IT	LT	LU	LV	MT	NL
auto_uni	0.1167	0.0455	0.0264	0.1219	0.1650	0.1011	0.1187	0.0217	0.1524	0.0312
auto_bi	0.0896	0.0710	0.0581	0.0896	0.1291	0.1301	0.1312	0.0413	0.1343	0.1695
auto	0.1009	0.0592	0.0449	0.1057	0.1452	0.1141	0.1252	0.0306	0.1431	0.1004
manual_old	0.0253		0.1598	0.4921				0.0989	0.1624	0.3799
manual_new			0.2172	0.1357						0.3276
manual	0.0253		0.1851	0.3139				0.0989	0.1624	0.3542
all	0.0979	0.0592	0.1396	0.1173	0.1452	0.1141	0.1252	0.0363	0.1440	0.2883

	PL	PT	RU	SE	SI	SK	UK
auto_uni	0.1348	0.1192	0.0129	0.0214	0.2085	0.0704	0.1110
auto_bi	0.1223	0.0538	0.0054	0.0367	0.1321	0.0458	0.1160
auto	0.1289	0.0835	0.0091	0.0316	0.1712	0.0610	0.1130
manual_old	0.128	0.1802	0.0899				0.2708
manual_new	0.5795						0.4376
manual	0.3537	0.1802	0.0899				0.3764
all	0.1401	0.1506	0.0450	0.0316	0.1712	0.0610	0.2365

6 Conclusion

The web is a natural reflection of the language diversity in the world, both in terms of web content as well as in terms of web users. Effective cross-language information retrieval (CLIR) techniques have clear potential for improving the search experience of such users. The WebCLEF track at CLEF 2006 attempts to realize some of this potential, by investigating known-item retrieval in a multilingual setting. Known-item retrieval is a typical search task on the web [3]. This year's track focused on mixed monolingual search, in which the topic set is a stream of known-item topics in various languages. This task was pioneered at WebCLEF 2005 [9]. The collection is based on the spidered content of web sites of European governments. This year's topic set covered all 27 primary domains in the collection, and contained both manually constructed search topics and automatically generated topics. Our main findings for the mixed-monolingual task are the following. First, the results over all topics show that current CLIR systems are quite effective. These systems retrieve, on average, the target page in the top ranks. This is particularly impressive when considering that the topics

of WebCLEF 2006 covered no less than 27 European primary domains. Second, when we break down the scores over the manually constructed and the generated topics, we see that the manually constructed topics result in higher performance. The manual topics consisted of both a set of newly constructed topics, and a selection of WebCLEF 2005 topics. For veteran participants, we can compare the scores over years, and we see progress for the old manual topics. The new manual topics (which were not available for training) confirm this progress.

Building a cross-lingual test collection is a complex endeavor. Information retrieval evaluation requires substantial manual effort by topic authors and relevance assessors. In a cross-lingual setting this is particularly difficult, since the language capabilities of topic authors should sufficiently reflect the linguistic diversity of the used document collection. Alternative proposals to traditional topics and relevance assessments, such as term relevance sets, still require human effort (albeit only a fraction) and linguistic capacities by the topic author.¹ This prompted us to experiment with techniques for automatically generating known-item search requests. The automatic construction of known-item topics has been applied earlier in a monolingual setting [2]. At WebCLEF 2006, two refined versions of the techniques were applied in a mixed-language setting. The general set-up of the the WebCLEF 2006 track can be viewed as an experiment with automatically constructing topics. Recall that the topic set contained both manual and automatic topics. This allows us to critically evaluate the performance on the automatic topics with the manual topics, although the comparison is not necessarily fair given that the manual and automatic subsets of topics differ both in number and in the domains they cover. Our general conclusion on the automatic topics is a mixed one: On the one hand, our results show that there are still some substantial differences between the automatic topics and manual topics, and it is clear that automatic topics cannot simply substitute manual topics. Yet on the other hand, the resulting scores on automatic topics give, at least, a solid indication of performance, and can hence be an attractive alternative in situations where manual topics are not readily available.

Acknowledgments. Thanks to *Universidad Complutense de Madrid* (UCM) for providing additional Spanish topics.

Krisztian Balog was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.065.120 and 612.000.-106. Jaap Kamps was supported by NWO under project numbers 612.066.302, 612.066.513, 639.072.601, and 640.001.501; and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.302, 612.069.006, 640.001.501, 640.002.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

¹ Recall that term relevance sets (T-rels) consisting of a set of terms likely to occur in relevant documents, and a set of irrelevant terms (especially disambiguation terms avoiding false-positives) [1].

References

- [1] Amitay, E., Carmel, D., Lempel, R., Soffer, A.: Scaling ir-system evaluation using term relevance sets. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 10–17. ACM Press, New York (2004)
- [2] Azzopardi, L., de Rijke, M.: Automatic construction of known-item finding test beds. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 603–604. ACM Press, New York (2006)
- [3] Broder, A.: A taxonomy of web search. *SIGIR Forum* 36(2), 3–10 (2002)
- [4] Eurobarometer. Europeans and their languages. Special Eurobarometer 243, European Commission, URL (2006), http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf
- [5] Gey, F.C., Kando, N., Peters, C.: Cross language information retrieval: a research roadmap. *SIGIR Forum* 36(2), 72–80 (2002)
- [6] Lemur: The Lemur toolkit for language modeling and information retrieval, URL (2005), <http://www.lmurproject.org/>
- [7] Miller, D.R., Leek, T., Schwartz, R.M.: A hidden Markov model information retrieval system. In: Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, US, pp. 214–221. ACM Press, New York (1999)
- [8] Sigurbjörnsson, B., Kamps, J., de Rijke, M.: EuroGOV: Engineering a multilingual Web corpus. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [9] Sigurbjörnsson, B., Kamps, J., de Rijke, M.: Overview of WebCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [10] Soboroff, I., Nicholas, C., Cahan, P.: Ranking retrieval systems without relevance judgments. In: *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 66–73. ACM Press, New York (2001)
- [11] Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 315–323. ACM Press, New York (1998)
- [12] WebCLEF. Cross-lingual web retrieval (2006), URL <http://ilps.science.uva.nl/WebCLEF/>

A Breakdown of Scores over Topic Types

We provide a breakdown of scores over the different topic types, both for the original topic set in Table 9 and for the new topic set in Table 10.

Table 9. Original topic set: MRR scores, for all runs submitted to WebCLEF 2006, by topic type. Best scoring run per team is in boldface.

RUN	ALL topics	AUTO			MANUAL		
		all	uni	bi	all	old	new
<i>buap</i>							
allpt40bi	0.0157	0.0040	0.0031	0.0049	0.0750	0.0810	0.0657
<i>depok</i>							
UI1DTA	0.0404	0.0234	0.0296	0.0173	0.1263	0.1099	0.1522
UI2DTF	0.0918	0.0466	0.0525	0.0406	0.3216	0.2611	0.4168
UI3DTAF	0.0253	0.0142	0.0116	0.0168	0.0819	0.0644	0.1094
UI4DTW	0.0116	0.0025	0.0020	0.0030	0.0583	0.0284	0.1053
<i>hildesheim</i>							
UHi1-5-10	0.0718	0.0242	0.0231	0.0253	0.3134	0.2550	0.4051
UHi510	0.0718	0.0242	0.0231	0.0253	0.3134	0.2550	0.4051
UHiBase	0.0795	0.0346	0.0328	0.0363	0.3076	0.2556	0.3893
UHiBrf1	0.0677	0.0220	0.0189	0.0251	0.3000	0.2485	0.3812
UHiBrf2	0.0676	0.0221	0.0188	0.0253	0.2989	0.2464	0.3816
UHiTitle	0.0724	0.0264	0.0245	0.0283	0.3061	0.2542	0.3876
UHiMu (<i>multilingual</i>)	—	—	—	—	0.2553	0.2146	0.3192
<i>hummingbird</i>							
humWC06	0.1133	0.0530	0.0572	0.0488	0.4194	0.3901	0.4657
humWC06dp	0.1209	0.0528	0.0555	0.0501	0.4664	0.4471	0.4967
humWC06dpc	0.1169	0.0472	0.0481	0.0464	0.4703	0.4553	0.4939
humWC06dpcD	0.1380	0.0704	0.0721	0.0687	0.4814	0.4633	0.5099
humWC06p	0.1180	0.0519	0.0556	0.0482	0.4538	0.4252	0.4988
<i>isla</i>							
baseline	0.1694	0.1253	0.1397	0.1110	0.3934	0.3391	0.4787
comb	0.1685	0.1208	0.1394	0.1021	0.4112	0.3578	0.4952
combmeta	0.1947	0.1505	0.1670	0.1341	0.4188	0.3603	0.5108
combNboost	0.1954	0.1586	0.1595	0.1576	0.3826	0.3148	0.4891
combPhrase	0.2001	0.1570	0.1639	0.1500	0.4190	0.3587	0.5138
<i>reina</i>							
usal_base	0.0100	0.0028	0.0044	0.0011	0.0468	0.0550	0.0340
usal_mix	0.0137	0.0038	0.0065	0.0011	0.0640	0.0747	0.0472
USAL_mix_hp	0.0139	0.0038	0.0065	0.0011	0.0655	0.0771	0.0472
usal_mix_hp	0.0139	0.0038	0.0065	0.0011	0.0655	0.0771	0.0472
usal_mix_hp_ok	0.0139	0.0038	0.0065	0.0011	0.0655	0.0771	0.0472
<i>rfia</i>							
DPSinDiac	0.0982	0.0721	0.0736	0.0706	0.2309	0.1808	0.3098
ERConDiac	0.1006	0.0771	0.0795	0.0746	0.2203	0.1693	0.3006
ERFinal	0.1021	0.0785	0.0803	0.0766	0.2220	0.1635	0.3140
ERSinDiac	0.1021	0.0785	0.0803	0.0766	0.2220	0.1635	0.3140
<i>ucm</i>							
webclef-run-all-2006-def-ok-2	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-2006-def-ok	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-2006-ok-conref	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-2006	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183
webclef-run-all-OK-definitivo	0.0870	0.0556	0.0578	0.0534	0.2461	0.2002	0.3183

Table 10. New topic set: MRR scores, for all runs submitted to WebCLEF 2006, by topic type. Best scoring run per team is in boldface.

RUN	ALL topics	AUTO			MANUAL		
		all	uni	bi	all	old	new
<i>buap</i>							
allpt40bi	0.0272	0.0080	0.0061	0.0099	0.0790	0.0863	0.0679
<i>depok</i>							
UI1DTA	0.0699	0.0465	0.0578	0.0348	0.1330	0.1171	0.1572
UI2DTF	0.1589	0.0923	0.1024	0.0819	0.3386	0.2783	0.4307
UI3DTAF	0.0439	0.0281	0.0226	0.0339	0.0862	0.0686	0.1130
UI4DTW	0.0202	0.0049	0.0038	0.0060	0.0613	0.0302	0.1088
<i>hildesheim</i>							
UHi1-5-10	0.1243	0.0480	0.0451	0.0510	0.3299	0.2717	0.4187
UHi510	0.1243	0.0480	0.0451	0.0510	0.3299	0.2717	0.4187
UHiBase	0.1376	0.0685	0.0640	0.0731	0.3238	0.2724	0.4023
UHiBrf1	0.1173	0.0436	0.0369	0.0505	0.3159	0.2648	0.3939
UHiBrf2	0.1171	0.0438	0.0367	0.0510	0.3147	0.2625	0.3943
UHiTitle	0.1254	0.0524	0.0479	0.0570	0.3222	0.2709	0.4005
UHiMu (<i>multilingual</i>)	–	–	–	–	0.2686	0.2286	0.3297
<i>hummingbird</i>							
humWC06	0.1962	0.1051	0.1116	0.0984	0.4416	0.4156	0.4812
humWC06dp	0.2092	0.1047	0.1084	0.1009	0.4910	0.4764	0.5132
humWC06dpc	0.2023	0.0937	0.0939	0.0935	0.4952	0.4852	0.5104
humWC06dpcD	0.2390	0.1396	0.1408	0.1384	0.5068	0.4936	0.5269
humWC06p	0.2044	0.1030	0.1086	0.0971	0.4777	0.4530	0.5154
<i>isla</i>							
baseline	0.2933	0.2485	0.2726	0.2237	0.4141	0.3614	0.4946
comb	0.2918	0.2394	0.2720	0.2058	0.4329	0.3812	0.5117
combmeta	0.3370	0.2985	0.3259	0.2701	0.4409	0.3839	0.5278
combNboost	0.3384	0.3145	0.3114	0.3176	0.4028	0.3355	0.5054
combPhrase	0.3464	0.3112	0.3199	0.3023	0.4411	0.3822	0.5310
<i>reina</i>							
usal_base	0.0174	0.0055	0.0087	0.0023	0.0493	0.0586	0.0351
usal_mix	0.0237	0.0075	0.0126	0.0022	0.0674	0.0796	0.0488
USAL_mix_hp	0.0241	0.0075	0.0126	0.0022	0.0689	0.0822	0.0488
usal_mix_hp	0.0241	0.0075	0.0126	0.0022	0.0689	0.0822	0.0488
usal_mix_hp_ok	0.0241	0.0075	0.0126	0.0022	0.0689	0.0822	0.0488
<i>rfia</i>							
DPSinDiac	0.1700	0.1429	0.1436	0.1422	0.2431	0.1926	0.3201
ERConDiac	0.1742	0.1528	0.1552	0.1503	0.2320	0.1804	0.3106
ERFinal	0.1768	0.1556	0.1568	0.1544	0.2337	0.1743	0.3244
ERSinDiac	0.1768	0.1556	0.1568	0.1544	0.2337	0.1743	0.3244
<i>ucm</i>							
webclef-run-all-2006-def-ok-2	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-2006-def-ok	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-2006-ok-conref	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-2006	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289
webclef-run-all-OK-definitivo	0.1505	0.1103	0.1128	0.1077	0.2591	0.2133	0.3289

Improving Web Pages Retrieval Using Combined Fields

Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez,
and Emilio Rodríguez

REINA Research Group, University of Salamanca, Spain
`reina@usal.es`
<http://reina.usal.es>

Abstract. This article describes the participation of the REINA Research Group of the University of Salamanca in WebCLEF 2006. This year we participated in the Monolingual Mixed Task in Spanish. The entire EuroGOV collection was processed to select all the pages in Spanish. All the pages with domain `.es` were also pre-selected. Our objective this year was to try pre-retrieval techniques of combining information fields or elements from web pages as well as the retrieval capability of these fields. In vector-based retrieval systems, the combining of terms coming from different sources can be achieved by operating on the frequency of the terms in the document using a weight scheme of $tf \times idf$. The BODY field is, of course, the most useful from the retrieval perspective, but the text of the backlinks brings considerable improvement. META fields or tags, however, contribute little to retrieval improvement.

1 Introduction

This article describes the participation of the REINA Research Group in WebCLEF 2006. The task we participated in this year was the Monolingual Mixed Task in Spanish. The basic document collection was the same as in 2005 [1], as were part of the queries. On this occasion we opted to broaden the document base to all those which, although pertaining to other domains, are written in Spanish.

Unfortunately, the page headers do not provide reliable information, either about the language or anything else; in many cases the headers are simply empty, while others have contents bearing little relation to reality. We thus had to resort to a language detector to select the pages in Spanish from other European domains. The detector selected was *TextCat* [2], a program based on the drawing up of patterns with *n-grams* for each possible language and the subsequent categorization of the text whose language is to be solved [3]. All the queries in Spanish were processed, including *named pages*, *home pages*, manually created ones and automatically created ones. All of them were treated in the same way.

This article is organized as follows: in the first section we describe the focus chosen to solve the task; in the following section we set out the problems and the options adopted for lexical analysis and extraction of terms from the pages.

Next, the information element fields that can be used in indexing and retrieval are discussed. This is followed by a description of the runs made and a discussion of the outcomes. We end with some conclusions.

2 Our Approach

The basic strategy used was: first, find the most relevant pages for each query and then according to the type of query, reorder the list of most relevant documents found. Finding the relevant pages for a given query can be approached by a conventional system of indexing and information retrieval, such as those based, for example, on the vector space model. However, there are more information elements on a web page than the text that users see in the window of their navigators. Even within that text we can find certain structures that will help, together with those other information elements, to improve retrieval.

For example, some of those elements cited are the **TITLE** field, some are **META** tags, backlink anchors, etc. Within the **BODY** field, which is what the navigator shows, we can differentiate parts that use different typographies, for example.

We thus have different sources of information that we must combine or fuse. Two basic fusion strategies have been proposed: pre-retrieval fusion and post-retrieval fusion [4]. The latter strategy is the one we used in WebCLEF 2005 and basically consists of building an index for each information element to be fused. Every time a query must be solved, it is executed against each of the indexes and the results obtained in each index are then fused together.

On this occasion we wanted to try out the pre-retrieval strategy, which consists of drawing up a single index with the terms of all the elements, but weighted in a different way. Once this single index has been made, the queries are executed normally against it. Naturally, in the making of this single index we may wish to value more or less the terms coming from such and such field or information element; this will allow us to use a combination which, if it is fine-tuned, should provide good retrieval results.

The application of a different valuation to the different components of the combination was easy in our case. Since for indexing and retrieval we use our software *Karpanta* [5], based on the well-known model of vector space, we only had to operate on the frequency of each term in each of the components of the combination.

In this case we applied an *ATU* weight scheme (*slope* = 0.2) [6], but the idea is similar for any weight scheme based on $tf \times idf$. We can apply a coefficient to tf as a function of what component of the combination the term appears on, always taking into account its frequency in the document and the *IDF*.

As to the second part of our strategy, reordering the list of documents or pages retrieved as a function of the type of query, we only considered the case of *home page* type queries.

To locate *home pages*, a simple strategy was followed. First, we manually detected the topics that could be of the *home page* type, although this is a strongly subjective assessment. Then for the results of these topics, a coefficient was

calculated based on two criteria: first, a simple heuristic based on the existence of certain expressions in the **TITLE** (*main page*, *welcome*, etc.). A similar heuristic was also applied to the name of the file (for example, *main.html*, *home.html*, etc.). The other criterion is the **URL** length, understood as **URL** path levels [7]: the fewer the levels the more probable that it is a home page [8].

The coefficient obtained is simply multiplied by the similarity of each page retrieved on each topic of the *home page* type, reordering the results of each retrieval.

2.1 Lexical Analysis and Term Extraction

The first operation, which must take place before any other, is the conversion of the web pages to plain text. This is necessary even to be able to determine the language of each page and whether to select it or not. Obtaining the plain text is not trivial and is not without problems.

As we stated before, one cannot take it on trust that standards are followed. It cannot even be guaranteed that the contents is **HTML**, even when the page begins with the `<html>` tag. So one may run directly into a binary code, or **PDFs** and the like. The same thing happens with **META** tags; even when they are present, they do not always offer the correct information.

In fact, many pages do not even contain text, so the first thing that has to be done is to determine the type of contents. The old and well-known **file** command, if fine-tuned, can help to determine the real contents of each page. It will also give us another important piece of data: the encoding system used. For pages in Spanish, this is usually **ISO 8859** or **UTF-8**; this information must be known in order to adequately process the special characters.

When the page turns out to be **HTML** the conversion to plain text is carried out with **w3m**. There are other converters, but after several trials the best results were obtained with **w3m**.

Once we have obtained the plain text, we can determine the language with **TextCat**. Documents in Spanish, as well as those pertaining to the **.es** domain will form part of our collection. From these, we must extract the terms and standardize them in some way. Basically, the characters are converted into small letters, and numbers, accents, spelling marks, and the like are eliminated. Stop words are eliminated and the terms are passed through an improved *s-stemming* [9].

2.2 Fields Used

There are diverse information elements or sources that can be considered on a web page. The base is the **BODY** field, obviously, but we can also use the **TITLE** field, which seems clearly descriptive, as well as several **META** tags usually employed to this end, especially those that contain *name="Description"* and *name="Keywords"*. However, as was already pointed out in [10], these fields are not present in a uniform way on all the pages. Many pages do not contain them, and others contain values placed automatically by the programs that have produced the page.

In addition, we can consider the terms with highlighted typography to be the most representative. The **H1**, **H2** tags or fields are an example. Unfortunately, the use of these tags is not uniform either and they lose ground in favour of the definition of fonts and specific sizes of characters, which is more difficult to process.

Another important element is the backlink anchors the pages receive. These anchors describe the page they are linked to in a more or less brief way; this description is important, because it is made by someone other than the author of the page we wish to index. It can be assumed that this description can add different terms to the ones used in the page itself; however, many of these anchors can refer to very specific parts of the page in question. Likewise, there are pages with many backlinks and many anchors, and others with few or none. And in any case, there is the matter of obtaining these anchors.

In our case, we processed the whole EuroGOV collection to obtain all the anchors and links to the pages in Spanish or those with the **.es** domain. Obviously there are more links to these pages outside EuroGOV, but we have no way of finding them.

Finally, we constructed the indexes with the following elements: **BODY**, **TITLE**, **meta-description**, **meta-keywords**, **meta-title**, **H1**, **H2**, **anchors**.

3 Executed Runs

With the topics and relevance estimations we carried out a training phase that allowed us to try out different combinations of elements in different proportions. We thus made three official runs and several unofficial ones. The first was based on an index constructed only with the body field and served a basis for comparison. The second was a combination of all the fields mentioned above. The third run was based on the same index as the second one, but with *Home Pages* boosting added, based on the size of the URLs and a simple heuristic, as mentioned above.

Table 1. Composition of the Indexes of the Official Runs

RUN 1	$BODY(fd \times 1)$
RUN 2	$RUN1 + ANCHOR(fd \times 1) + TITLE(fd \times 1.5) + META-DESC(fd \times 1.5) + META-TITLE(fd \times 1) + META-KEY(fd \times 0.5) + H1(fd \times 0.8) + H2(fd \times 0.8)$
RUN 3	Same as RUN 2 + <i>Home Pages</i> boosting

The outcomes of these official runs clearly confirm the advantage of using the additional information elements. Of these, the backlink **anchors** seem to be the most useful.

We also made several unofficial runs that endorse the official results. Each run was executed against an index drawn up with the terms of each information field or element considered ($fd \times 1$). Each run worked with only one of these fields,

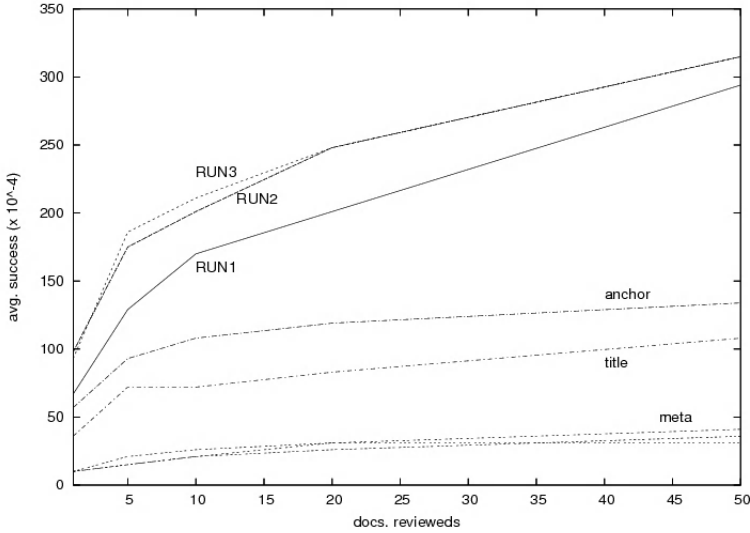


Fig. 1. Results of the Official and Unofficial Runs

without using the terms that appear in the body. The H1 and H2 fields are not shown in the figure since their results were extremely low.

As was to be expected, each field separately yielded worse results than **BODY**, which is normal since the **BODY** is where the visual text of the page is located. But, apart from the **BODY**, the field with the most retrieval power is the backlinks **anchor**, even though in many cases these **anchors** are very short. It does seem, however, that, as stated before, the descriptions that others make of a page are quite useful for retrieval purposes.

This is closely followed by the **TITLE** field, which was also expected. However, the fields based on **META** tags yielded poor outcomes, far behind **anchors** and **TITLE**, despite being fields expressly orientated to retrieval. Although there was little difference between the outcomes of the three **META** tags observed (**title**, **description** and **keywords**), it is the last of these which curiously enough worked the least well of all three.

4 Conclusions

We have described our participation in WebCLEF 2006, the strategy we adopted, the experiments carried out and the results obtained. The use of information elements or fields in addition to the text or **BODY** of the pages improved retrieval outcomes. One way of using these fields is to construct a single index with the terms that appear in them, together with the words that appear in the **BODY** of the page. This requires the terms of each field to be assessed differently, and this assessment must be open to empirical adjustment.

Of all the fields, the one formed by the backlinks **anchors** that each page receives seems to be the most effective. The **TITLE** field also contributes notably to improving retrieval. The contents of the **META** tags, however, seems to be of limited use from the point of view of retrieval.

Acknowledgements

This research was partially funded by the Regional Government of Castile & Leon (Spain) with Project SA089/04.

References

1. Sigurbjörnsson, B., Kamps, J., Rijke, M.d.: Overview of webclef (2005) [11]
2. Noord, G.v.: (Texcat language guesser)
3. Cavnar, W.B., Trenkle, J.M.: N-gram-based text categorization. In: Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, April 11-13, 1994, pp. 161–175 (1994)
4. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O., Goharian, N.: On fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology (JASIST)* 55(10), 859–868 (2004)
5. Figuerola, C.G., Alonso Berrocal, J.L.A., Zazo Rodríguez, Á.F., Rodríguez Vázquez de Aldana, E.: Herramientas para la investigación en recuperación de información: Karpanta, un motor de búsqueda experimental. *Scire* 10(2), 51–62 (2004)
6. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, (Special Issue of the SIGIR Forum) August 18–22, 1996, pp. 21–29. ACM Press, New York (1996)
7. Kraaij, W., Westerveld, T., Hiemstra, D.: The importance of prior probabilities for entry page search. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 27–34. ACM Press, New York (2002)
8. Tomlinson, S.: Robust, web anf terabyte retrieval with hummingbird searchserver at trec 2004. In: The Thirteenth Text Retrieval Conference (TREC 2002), pp. 261–500. NIST Special Publication (2004)
9. Figuerola, C.G., Zazo, Á.F., Rodríguez Vázquez de Aldana, E., Alonso Berrocal, J.L.: La recuperación de información en español y la normalización de términos. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* 8(22), 135–145 (2004)
10. Figuerola, C.G., Alonso Berrocal, J.L., Zazo Rodríguez, Á.F., Rodríguez, E.: REINA at the WebCLEF task: Combining evidences and link analysis [11]
11. Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.): Results of the CLEF 2005 Cross-Language System Evaluation Campaign. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)

A Penalisation-Based Ranking Approach for the Mixed Monolingual Task of WebCLEF 2006*

David Pinto^{1,2}, Paolo Rosso¹, and Ernesto Jiménez³

¹ Department of Information Systems and Computation, UPV, Spain

² Faculty of Computer Science, BUAP, Mexico

³ School of Applied Computer Science, UPV, Spain

{dpinto, prosson}@dsic.upv.es, erjica@ei.upv.es

Abstract. This paper presents an approach of a cross-lingual information retrieval which uses a ranking method based on a penalisation version of the Jaccard formula. The obtained results after the submission of a set of runs to the WebCLEF 2006 have shown that this simple ranking formula may be used in a cross-lingual environment. A comparison with runs submitted by other teams ranks us in a third place by using all the topics. A fourth place is obtained with our best overall results by using only the new topic set, and a second place was got by using only the automatic topics of the new topic set. An exact comparison with the rest of the participants is in fact difficult to obtain and, therefore, we consider that further detailed analysis of the components should be done in order to determine the best components of the proposed system.

Keywords: Retrieval models, Mixed-Monolingual search process.

1 Introduction

The current commercial search engines, such as Google and Yahoo, provide only monolingual information retrieval and, therefore, forums dedicated to the analysis and evaluation of information retrieval systems in a cross-lingual environment, such as WebCLEF [1], are needed. Since its first edition in 2005, the WebCLEF concern has been to deal with the EuroGOV corpus, which consists in a crawl of European governmental sites from approximately 27 different Internet domains [5,3]. The aim of this work is to evaluate a new similarity measure in the Mixed-Monolingual task of WebCLEF evaluation forum. The introduced formula is a variation of the Jaccard coefficient with a penalisation factor.

In the next section we will describe the way we have processed this corpus in order to obtain the index terms. Section 3 explains the model we have implemented, whereas its evaluation is presented in Section 4. Finally, a discussion of our participation and the obtained results in this evaluation forum are given.

* This work was partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

2 Dataset Preprocessing

We have written two scripts in order to obtain the index terms of the EuroGOV corpus. The first script uses regular expressions for excluding all the information which is enclosed by the characters < and >. This script obtains very good results, but it is very slow and, therefore, we decided to use it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE). We have wrote another script based on the *html* syntax for obtaining all the terms considered valuable for indexing, which speeded up our indexing process. Unfortunately, major web pages do not strictly observe the *html* syntax and, therefore, we missed important information from those documents. We have implemented different methods for detecting the charset codification of those webpages that are not in UTF-8. We have observed that the charset codification detection is one the most difficult problem in the preprocessing step. Finally, we eliminated stop words for each language (except Greek) and punctuation symbols. The same process was applied to the queries.

3 The Penalisation-Based Ranking Approach

Nowadays, different information retrieval models are reported in literature [4] [2]. In this work, we have used a variation of the boolean model with ranking based in the Jaccard similarity formula. We named this variation “Jaccard with penalisation”, because it punishes the ranking score taking into account the number of terms that a query Q_i really matches when it is compared with a document D_j of the collection. The formula used is presented as follows:

$$Score(Q_i, D_j) = \frac{|D_j| \cap |Q_i|}{|D_j| \cup |Q_i|} - \left(1 - \frac{|D_j| \cap |Q_i|}{|Q_i|}\right) \quad (1)$$

As can be seen, the first component of this formula is the typical Jaccard approximation. The evaluation of this formula is quite fast, and allows its implementation in real situations. The obtained results by using this approach are presented in the next section.

4 Experimental Results

We have submitted three different runs in order to experiment with the use of diacritics in the corpus preprocessing step. We have renamed all the runs in this paper (cursive), with respect to the names reported in [1] (bold face) as follows: *WithoutDiac* (**ERFinal**), *WithDiac* (**ERConDiac**), *CDWithoutDiac* (**DPSinDiac**).

Table 1 shows the results obtained with each of the three different approximations submitted. The *WithoutDiac* run eliminates all diacritics in both, the

corpus and the topics, whereas the *WithDiac* run only supresses the diacritics in the corpus. We may observe an expected reduction of the Mean Reciprocal Rank (MRR), but it does not differ significantly from the first run. This is clearly derived from the amount of diacritics introduced in the evaluation topics set, which is not very high. An analysis of the queries in real situations may be interesting in order to determine whether the topics set is realistic. The last run (*CDWithoutDiac*) eliminates diacritization in both, the topics and corpus; besides, it tries a charset detection for each indexed document.

Table 1. Evaluation of each run submitted

Team	Average Success at					50	MRR over 1939
	Run	1	5	10	20		
rfia	<i>WithoutDiac</i>	0,0665	0,1423	0,1769	0,2192	0,2625	0,1021
rfia	<i>WithDiac</i>	0,0665	0,1372	0,1717	0,2130	0,2568	0,1006
rfia	<i>CDWithoutDiac</i>	0,0665	0,1310	0,1681	0,1996	0,2470	0,0982

Table 2 shows a summary of all the best participant runs submitted to the mixed monolingual task of WebCLEF 2006. The Mean Reciprocal Rank (MRR) scores are reported for both the original and the new topic set. The first column indicates the name that each team had in the evaluation forum, whereas the second column indicates the name of their best run. The scores shown in that table rank us in a third place.

Table 2. Best runs for each WebCLEF 2006 participant in the mixed monolingual task

Team Name	Run Name	MRR for the original topic set	MRR for the new topic set
isla	CombPhrase	0.2001	0.3464
hummingbird	humWC06dpcD	0.1380	0.2390
rfia	WithoutDiac (ERFinal)	0.1021	0.1768
depok	UI2DTF	0.0918	0.1589
ucm	webclef-run-all-2006	0.0870	0.1505
hildesheim	UHiBase	0.0795	0.1376
buap	allpt40bi	0.0157	0.0272
reina	USAL_mix_hp	0.0139	0.0241

In Table 3(a) we can see the best overall results by using only the new topic set. Here we have obtained a fourth place, according to the average among the automatic and the manual topic scores. Whereas, in Table 3(b) we may observe the results by using only the new automatic generated topics. Our second place shows that the penalisation-based ranking is working well for the task proposed in this evaluation forum.

Table 3. Best overall runs for each WebCLEF 2006 participant by using the new topic set with: (a) all topics, and (b) only the automatically generated topics

Team Name	all	auto	manual	average	auto-uni	auto-bi
isla	0.3464	0.3145	0.4411	0.3778	0.3114	0.3176
hummingbird	0.2390	0.1396	0.5068	0.3232	0.1408	0.1384
depok	0.1589	0.0923	0.3386	0.2154	0.1024	0.0819
rfia	0.1768	0.1556	0.2431	0.1993	0.1568	0.1544
hildesheim	0.1376	0.0685	0.3299	0.1992	0.0640	0.0731
ucm	0.1505	0.1103	0.2591	0.1847	0.1128	0.1077
buap	0.0272	0.0080	0.0790	0.0435	0.0061	0.0099
reina	0.0241	0.0075	0.0689	0.0382	0.0126	0.0022
	(a)				(b)	

5 Discussion

We have proposed a new approach for the ranking formula in an information retrieval system which is based on the Jaccard formula, but with a penalisation factor. After evaluating this approach in the approximately 75% of queries from the WebCLEF evaluation forum, we have obtained the third place in the overall results, among eight participant teams. In the particular case of the new topic set, we obtained fourth place in overall (all topics), and second place using automatically generated topics. An evaluation of the use of diacritization in the task has shown that results are not significantly different, which may be suggesting that the set of queries provided for the evaluation does not have a high number of diacritics. Further investigation would determine whether this behaviour is realistic or must be tuned in further evaluations. Finally, we consider that the system proposed may be improved by taking into account a better understanding of the preprocessing phase.

References

1. Balog, K., Azzopardi, L., Kamps, J., de Rijke, M.: Overview of WebCLEF 2006. In this Volume (2007)
2. Kraaij, W., Simard, M., Nie, J.Y.: Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics* 29(3), 381–419 (2003)
3. Pinto, D., Jiménez-Salazar, H., Rosso, P.: BUAP-UPV TPIRS: A System for Document Indexing Reduction on WebCLEF. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)
4. Salton, G.: *Automatic Text Processing*. Addison-Wesley, Reading (1989)
5. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: EuroGOV: Engineering a Multilingual Web Corpus. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, Springer, Heidelberg (2006)

Index Combinations and Query Reformulations for Mixed Monolingual Web Retrieval

Krisztian Balog and Maarten de Rijke

ISLA, University of Amsterdam
kbalog,mdr@science.uva.nl

Abstract. We examine the effectiveness on the multilingual WebCLEF 2006 test set of light-weight methods that have proved successful in other web retrieval settings: combinations of document representations on the one hand and query reformulation techniques on the other.

We investigate a range of approaches to crosslingual web retrieval using the test suite of the mixed monolingual CLEF 2006 WebCLEF track, featuring a stream of known-item topics in various languages. The topics are a mixture of manual (human generated) and automatically generated topics. We examine the robustness of well-known web retrieval techniques on this test set: compact document representations (titles or incoming anchor-texts), and query reformulation techniques. In Section 1 we describe our retrieval system as well as the approaches we applied. In Section 2 we describe our experiments, while the results are detailed in Section 3. We conclude in Section 4. For details on the WebCLEF collection and on the topics used we refer to [1].

1 System Description

Our retrieval system is based on the Lucene engine [5]. For ranking, we used the default similarity measure of Lucene, i.e., for a collection D , document d and query q containing terms t_i :

$$\text{sim}(q, d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where

$$tf_{t,X} = \sqrt{\text{freq}(t, X)} \quad idf_t = 1 + \log \frac{|D|}{\text{freq}(t, D)},$$
$$norm_d = \sqrt{|d|} \quad coord_{q,d} = \frac{|q \cap d|}{|q|}, \quad \text{and} \quad norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t^2}.$$

We did not apply any stemming nor did we use a stopword list. We applied case-folding and normalized marked characters to their unmarked counterparts, i.e., mapping á to a, ö to o, æ to ae, î to i, etc. The only language specific processing we did was a transformation of the multiple Russian and Greek encodings into an ASCII transliteration.

We extracted the full text from the documents, together with the title and anchor fields, and created three separate indexes: *content* (index of the full document text), *title* (index of all <title> fields), and *anchors* (index of all incoming anchor-texts). We created three base runs using the separate indexes and evaluated the base runs using the WebCLEF 2005 topics. Based on the outcomes we decided to use only the *content* and *title* indexes.

1.1 Run Combinations

We experimented with combinations of *content* and *title* runs, using standard run combination methods Fox and Shaw [2]: **combMAX** (take the maximal similarity score of individual runs); **combMIN** (take the minimal similarity score of individual runs); **combSUM** (take the sum of the similarity scores of individual runs); **combANZ** (take the sum of the similarity scores of individual runs, and divide by the number of non-zero entries); **combMNZ** (take the sum of the similarity scores of individual runs, and multiply by the number of non-zero entries); and **combMED** (take the median similarity score of individual runs).

Fox and Shaw [2] found **combSUM** to be the best performing combination method. Kamps and de Rijke [3] conducted extensive experiments with the Fox and Shaw combination rules for nine European languages, and demonstrated that combination can lead to significant improvements. Moreover, they showed that the effectiveness of combining retrieval strategies differs between English and other European languages. In Kamps and de Rijke [3], **combSUM** emerges as the best combination rule, confirming Fox and Shaw's findings.

Similarity score distributions may differ radically across runs. Therefore, we apply combination methods to normalized similarity scores only. We follow Lee [4] and normalize retrieval scores into a $[0, 1]$ range, using the minimal and maximal similarity scores (min-max normalization): $s' = (s - \min)/(max - \min)$, where s denotes the original retrieval score, and $\min$ (max) is the minimal (maximal) score over all topics in the run.

For the WebCLEF 2005 topics the best performance was achieved using the **combSUM** combination rule, which is in conjunction with the findings in [2, 3], therefore we used that method for the experiments on the WebCLEF 2006 topics.

1.2 Query Reformulation

In addition to our run combination experiments, we conducted experiments to measure the effect of phrase and query operations. We tested query-rewrite heuristics using phrases and word n -grams.

As to phrases, we simply added the topic terms as a phrase to the topic. For example, for the topic WC0601907, with title "diana memorial fountain", the query becomes: *diana memorial fountain* "diana memorial fountain". Our intuition is that rewarding documents that contain the whole topic as a phrase, not only the individual terms, would be beneficial to retrieval performance.

In our approach to n -grams, every word n -gram from the query is added to the query as a phrase with weight n . This means that longer phrases get a bigger boost. Using the previous topic as an example, the query becomes: *diana*

memorial fountain “*diana memorial*”² “*memorial fountain*”² “*diana memorial fountain*”³, where the number in the upper index denotes the weight attached to the phrase (the weight of the individual terms is 1).

2 Runs

To assess the effectiveness of index combinations and query reformulations, we created five runs using the WebCLEF 2006 mixed monolingual topics: *Baseline* (base run using the *content* only index); *Comb* (combination of the *content* and *title* runs, using the *combSUM* method); *CombMeta* (like the *Comb* run, but using the *target domain* meta field; we restrict our search to the corresponding domain); *CombPhrase* (the *CombMeta* run, but using the *Phrase* query reformulation technique); *CombNboost* (the *CombMeta* run, using the *N-grams* query reformulation technique).

3 Results

Table 1 lists our results in terms of mean reciprocal rank. Runs are listed along the left-hand side, while the labels indicate either all topics (*all*) or various subsets: automatically generated (*auto*)—further subdivided into automatically generated using the unigram generator (*auto-u*) and automatically generated using the bigram generator (*auto-b*)—and manual (*manual*), which is further subdivided into *new* manual topics and *old* manual topics; see [1] for details.

Significance testing was done using the two-tailed Wilcoxon Matched-pair Signed-Ranks Test, where we look for improvements at a significance level of 0.05 (¹), 0.001 (²), and 0.0001 (³). Significant differences noted in Table 1 are with respect to the **Baseline** run.

Combining the *content* and *title* runs (**Comb**) increased performance only for the manual topics. The use of the title tag does not help when the topics are automatically generated. Instead of employing a language detection method, we simply used the target domain meta field. The **CombMeta** run improved the retrieval performance significantly for all subsets of topics. Our query reformulation techniques show mixed, but promising results. The best overall score was achieved when the topic, as a phrase, was added to the query (**CombPhrase**). The comparison of **CombPhrase** vs **CombMeta** reveals that they achieved similar scores for all subsets of topics, except for the automatic topics using the bigram generator, where the query reformulation technique was clearly beneficial.

Table 1. Submission results (Mean Reciprocal Rank)

runID	all	auto	auto-u	auto-b	manual	man-n	man-o
Baseline	0.1694	0.1253	0.1397	0.1110	0.3934	0.4787	0.3391
Comb	0.1685 ¹	0.1208 ³	0.1394 ³	0.1021	0.4112	0.4952	0.3578
CombMeta	0.1947 ³	0.1505 ³	0.1670³	0.1341 ³	0.4188 ³	0.5108 ¹	0.3603¹
CombPhrase	0.2001³	0.1570 ³	0.1639 ³	0.1500 ³	0.4190	0.5138	0.3587
CombNboost	0.1954 ³	0.1586³	0.1595 ³	0.1576³	0.3826	0.4891	0.3148

The n -gram query reformulation technique (**CombNboost**) further improved the results of the *auto-b* topics, but hurt accuracy on other subsets, especially on the manual topics. The **CombPhrase** run shows that even a very simple query reformulation technique can be used to improve retrieval scores.

Comparing the various subsets of topics, we see that the automatic topics are more difficult than the manual ones. Also, the new manual topics seem to be more appropriate for known-item search than the old manual topics. There is a clear ranking among the various subsets of topics, and this ranking is independent from the applied methods: $man - n \gg man - o \gg auto - u \gg auto - b$.

4 Conclusions

We investigated the effectiveness of known web retrieval techniques in the mixed monolingual setting. The only language-specific processing we applied was the transformation of various encodings into an ASCII transliteration. We found that the standard combination **combSUM** combination using Min-Max normalization is effective only for human generated topics; using the title field did not improve performance when the topics are automatically generated. Significant improvements (+15% MRR) were achieved by using the target domain meta field. Furthermore, even simple query reformulation methods can improve retrieval performance, but not consistently across all subsets of topics. Although it may be too early to talk about a solved problem, effective and robust web retrieval techniques seem to carry over to the mixed monolingual setting.

Acknowledgments

Krisztian Balog was supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 600.-065.-120 and 612.000.106. Maarten de Rijke was supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.-065.-120, 612-13-001, 612.000.106, 612.-066.302, 612.069.006, 640.001.501, and 640.002.501.

References

- [1] Balog, K., Azzopardi, L., Kamps, J., de Rijke, M.: Overview of WebCLEF 2006. In: This Volume (2007)
- [2] Fox, E., Shaw, J.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2), pp. 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215 (1994)
- [3] Kamps, J., de Rijke, M.: The effectiveness of combining information retrieval strategies for European languages. In: Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 1073–1077. ACM Press, New York (2004)
- [4] Lee, J.H.: Combining multiple evidence from different properties of weighting schemes. In: SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 180–188. ACM Press, New York (1995)
- [5] Lucene. The Lucene search engine (2005), <http://lucene.apache.org/>

Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for WebCLEF 2006 at the University of Hildesheim

Ben Heuwing, Thomas Mandl, and Robert Strötgen

Universität Hildesheim, Information Science
Marienburger Platz 22, D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. Experiments with the analysis and extraction of the HTML structure of web documents were carried out for WebCLEF 2006. In addition, blind relevance feedback was applied. As for WebCLEF 2005, a language independent indexing strategy was pursued. We experimented with HTML title, H1 element and other elements emphasizing text. Our index contained title and H1, emphasized elements, full and partial content. The best results with the WebCLEF 2005 topics were achieved with a strong weight on the title-element and a very small weight on emphasized text leading to a marginal improvement over the best post submission runs for the mixed-monolingual task at WebCLEF 2005. For the WebCLEF 2006 topics, improved results were achieved for manually generated topics. The best performance for manual topics for WebCLEF 2006 was achieved with a strong weight on both HTML title as well as H1 elements, and a decreased weight for the other elements. Blind relevance feedback could not yet improve the results.

1 Introduction

The participation of the University of Hildesheim was based on the experience gained during WebCLEF 2005 [4]. For the multilingual task, our approach had shown competitive results in WebCLEF 2005. This year our efforts were centered on the mixed-monolingual task [1].

2 System Description

The system developed for WebCLEF 2005 was built on top of Apache Lucene. A language independent approach was followed. All documents were indexed in one multilingual index. As a consequence, no language dependent stemming algorithms could be applied. Indexing strategies were refined this year based on the structure of the HTML files. At WebCLEF 2005, retrieval based on the HTML title element proved to be effective for multilingual web retrieval. The titles might lead to further improvement. Some titles might be of low quality and could be eliminated. This hypothesis was based on the observation that many web pages have the title *no title* or similar phrases in other languages. In order to identify these phrases and assemble a

stop title list which would not be indexed, the frequency of title phrases was assessed. Surprisingly, the hypothesis did not hold for the EuroGOV corpus. Some titles do occur often, but they contain valuable text and should not be eliminated. As a consequence, the stopword list was extended with the most frequent title words.

The first order headline (H1) element was identified and added to the title in order to be indexed conjointly. A set of other HTML elements which emphasize text (H{1-6}, strong, b, em, bold, i) was identified and joined to form one indexing field. As in WebCLEF 2005, both full and partial content were indexed. Instead of choosing the first characters for the partial content cutoff, we adopted a refined strategy. The most important and discriminating content for a web page is often not at the beginning of the HTML code. In many cases, the beginning of the code contains navigation elements and menus which are stable for a whole site [2]. Consequently, 50 tokens from the middle of the HTML code were selected as the partial content. A more elaborated strategy was not adopted in order to not compromise the efficiency. Blind relevance feedback (BRF) is an efficient optimization strategy for ad-hoc retrieval. We adopted a BRF implementation for ad-hoc retrieval [3] and integrated it into the web retrieval scenario. BRF was realized for all index fields except for the full content due to hardware limitations. In addition, a domain filter was implemented which takes advantage of the meta-data provided with the topics.

3 Experimental Results

The parameters were optimized based on WebCLEF 2005 topics. Based on the results, runs with the parameters shown in Table 1 were submitted.

Table 1. Parameters for the submitted Runs

Run	Weights
<i>UHiBase</i>	content^1 emphasised^0.1 title^20
<i>UHiTitle</i>	content^1 emphasised^1 title^20
<i>UHi1-5-10</i>	content^1 emphasised^5 title^10
<i>UHiBrf1</i>	content^1 emphasised^1 title^20 BRF (weight of expanded query: 1)
<i>UHiBrf2</i>	content^1 emphasised^1 title^20 BRF (weight of expanded query: 0.5)
<i>UHiMu</i>	(multilingual) content^1 emphasised^1 title^20 – translated query^10

Results of the runs for the topics of WebCLEF 2005 are shown in Table 2. Compared to the best post submission experiments of WebCLEF 2005, it can be seen that the result of the multilingual experiments could be improved from 0.2117 to 0.2443. The top performing result submitted by the University of Hildesheim had been 0.1370.

For the mixed monolingual task, the performance was improved (MRR from 0.2377 to 0.2819) without reaching the performance levels of the best participants in 2005. The average success rate at position ten was improved from 0.235 to 0.4168.

These indexing strategies did improve the results overall, however, they did not lead to competitive results for the mixed monolingual task.

Table 2. Results of experiments (WebCLEF 2005 topics)

Run	UHiBase	UHiTitle	UHi1-5-10	UHiBrf1	UHiBrf2	UHiMu
MRR	0.2819	0.2807	0.2814	0.2731	0.2771	0.2443
Avg success at 10	0.4168	0.4132	0.4186	0.3949	0.4040	0.3656

As a base-run, the run which showed the best results in the experiments with the WebCLEF 2005 topics was chosen. The queries were weighted, the weight of the *emphasized*-field actually being decreased as preliminary results had shown that this leads to slightly better results. Additionally, a similar run with a heavily weighted *title*-field (*UHiTitle*) and one with moderate weights (*emphasised*⁵ and *title*¹⁰) were submitted. In all runs, the full-content field was used for search, as it leads to significantly better results than the partial-content field. To test the effect of BRF, two runs with different weights on the expanded query were generated. The performance of these runs was slightly below that of the - in all other aspect equivalent - *UHiTitle*-Run. According to these findings, BRF has not shown a positive effect on retrieval quality so far. To test the improvements in the multilingual context, a non-official run for the multilingual task was submitted.

The improvements achieved are partly a result of the use of meta-data, restricting search to the target-domain. This improved the position of relevant documents in the result list. Not using the filter, the *UHiTitle*-Run has a lower MRR of 0.2552, but the 'average success at ten'-rate of 0.3784 still shows an improvement, probably due to the exhaustively indexed corpus as well as the optimized weighting of the fields.

Table 3. Results WebCLEF 2006

	<i>all topics</i>		<i>manually generated topics</i>	
	MRR	Average success at 10	MRR	Average success at 10
UHiBase	0.0795	0.1377	0.3076	0.4451
UHiTitle	0.0724	0.1253	0.3061	0.4420
UHi1-5-10	0.0718	0.1233	0.3134	0.4577
UHiBrf1	0.0677	0.1104	0.3000	0.4295
UHiBrf2	0.0676	0.1124	0.2989	0.4295
UHiMu	0.0489	0.0758	0.2553	0.3824

The difference between the results of the different topic-types is considerable (Table 3). While on manually generated topics (319 of 1939) the runs performed as it was to be expected from the experimental results, the performance on automatically generated topics (1620 of 1939) was poor. For the manual topics, the *UHi1-5-10-Run* produced the best results with an MRR score of 0.3134 and an 'average success at 10'-rate of 0.4577. A relevant document was five times more likely to appear within the first ten hits as a result of a manual topic than as a result of an automatically generated one. Differences were also found between the results of the new manually created topics (124) and those of the old topics (195) of 2005.

Considering only the manually created topics, the results of WebCLEF 2006 show the improvements that were to be expected from the previous experiments. The results were even slightly better than those of the experiments with the WebCLEF 2005

topics. The best performance was achieved with a strong weight on HTML title and H1 elements, a moderate weight for the other elements extracted and without BRF. To assess the effects of weighting on the different fields, further experiments were carried out after submissions. Some of the results are shown in Table 4.

Table 4. Weights and Results of Post Submission Runs for WebCLEF 2005 topics

Content	Title	Emphasized	MRR	success@50
1	10	5	0.2814	0.5229
1	20	1	0.2806	0.5192
1	1	1	0.2842	0.5649
1	20	0.1	0.2819	0.521
1	20	0.001	0.2873	0.5229
1	0.00001	0.00001	0.3221	0.6453
1	0.00001	0.000001	0.3198	0.6453
1	0	0	0.2842	0.6325
0	1	0	0.2576	0.4388

A very small weight on the emphasized text resulted in an improvement compared to the submitted run. This may be due to the fact that the emphasized text was indexed both in its own field as well as the content and the length normalization between fields implemented by Lucene.

4 Conclusion and Outlook

For the second web track participation at WebCLEF we intended to tune our system and to index several fields. Different weights for fields lead to better results while blind relevance feedback (BRF) did not. In future experiments, we intend to test different weighting strategies for BRF, quality metrics and typical features of relevant pages [5].

References

1. Balog, K., Azzopardi, L., Kamps, J., Rijke, M.: Overview of WebCLEF In: this volume (2006)
2. Chen, L., Ye, S., Li, X.: Template Detection for Large Scale Search Engines. In: Proc ACM Symposium on Applied Computing, pp. 1094–1098. ACM Press, New York (2006)
3. Hackl, R., Mandl, T., Womser-Hacker, C.: Mono- and Cross-Lingual Retrieval Experiments at the University of Hildesheim. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 165–169. Springer, Heidelberg (2005)
4. Jensen, N., Hackl, R., Mandl, T., Strötgen, R.: Web Retrieval Experiments with the EuroGOV Corpus at the University of Hildesheim. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 837–845. Springer, Heidelberg (2006)
5. Kamps, J.: Web-centric language models. In: Proc. 14th ACM CIKM '05, pp. 307–308. ACM, New York (2005)

Vocabulary Reduction and Text Enrichment at WebCLEF*

Franco Rojas¹, Héctor Jiménez-Salazar², and David Pinto^{1,3}

¹ Faculty of Computer Science, BUAP, Mexico

² Department of Information Technologies, UAM, México

³ Department of Information Systems and Computation, UPV, Spain
{frrlb99, hgimenezs, davideduardopinto}@gmail.com

Abstract. Nowadays, cross-lingual Information Retrieval (IR) is one of the greatest challenges to deal with. Besides, one of the most important issues in IR consists of the corpus vocabulary reduction. In real situations some methods of IR such as the well-known vector space model, it is necessary to reduce the term space. In this work, we have considered a vocabulary reduction process based on the selection of mid-frequency terms. Our approach enhances precision, but in order to obtain a better recall, we have conducted an enrichment process based on the addition of co-occurrence terms. By using this approach, we have obtained an improvement of 40%, using the BiEnEs topics of the WebCLEF 2005 task. The obtained results in the current mixed monolingual task of the WebCLEF 2006 have shown that the text enrichment must be done before the vocabulary reduction process in order to get the best performance.

Keywords: text representation, information retrieval, multilingual information.

1 Introduction

The multilingual information published in Internet has experimented a big explosion in the last years. This fact lead us to develop novel techniques to deal with the current cross-language feature of the World Wide Web (WWW). A common language scenario in which a user may be interested in information which is in a different language than his/her own native language would be when this user has some comprehension ability for a given language but she/he is not sufficiently proficient to confidently specify a search request in that language. Thus, a search system that can deal with this problem should be of a high benefit. The WWW is a natural setting for cross-lingual information retrieval and the European Union is a typical example of a multilingual scenario, where multiple users have to deal with information published in several languages. Therefore, evaluation environments for cross-lingual information retrieval systems are needed.

* This work was partially supported by FCC-BUAP, DCCD-Cuajimalpa and the BUAP-701 PROMEP/103.5/05/1536 grant.

The Cross-Language Evaluation Forum (CLEF) promotes the evaluation of cross-lingual information retrieval systems for different types of data [4]. WebCLEF is a particular task for the evaluation of such systems that deals with information on the Web [7]. A detailed discussion of the teams participation and the specific characteristics of the task proposed in the current WebCLEF may be found in [2]. In fact, they have proposed mainly one task for the evaluation of cross-lingual search engines: the Mixed Monolingual task. Thus, in this paper we are reporting the obtained results after the submission of one run to this evaluation forum.

We have used a text reduction with an enrichment process and, therefore, we organized this document in four sections. The next section describes the components of our search engine. In Section 3 a discussion of the corpus preprocessing as well as the obtained evaluation results are presented. Finally a conclusion of findings are given.

2 Description of the Search Engine

The aim of this approach was to determine the behaviour of document indexing reduction in a cross language information retrieval environment. We have used a boolean model with the Jaccard similarity formula in order to rank the obtained documents related to some query. In order to reduce the terms of every document treated, we have applied a mid-frequency terms based technique, named Transition Point, which is described as follows.

2.1 The Transition Point Technique

The Transition Point (TP) is a frequency value that splits the vocabulary of a text into two sets of terms (low and high frequency). This technique is based on the Zipf Law of Word Occurrences [10] and also on the refined studies of Booth [1], as well as of Urbizagástegui [9]. These studies are meant to demonstrate that mid-frequency terms, of a text T , are closely related to the conceptual content of T . Therefore, it is possible to establish the hypothesis that terms closer to TP can be used as index terms of T . A typical formula used to obtain this value is: $TP = (\sqrt{8 * I_1 + 1} - 1)/2$, where I_1 represents the number of words with frequency equal to 1; see [5] [9].

Alternatively, TP can be localized by identifying the lowest frequency (from the highest frequencies) that it is not repeated in the text; this characteristic comes from the properties of the Booth's law of low frequency words [1]. In our experiments we have used this approach.

Let us consider a frequency-sorted vocabulary of a document; i.e., $V_{TP} = [(t_1, f_1), \dots, (t_n, f_n)]$, with $f_i \geq f_{i+1}$, then $TP = f_{i-1}$, iff $f_i = f_{i+1}$. The most important words are those nearest to the TP, i.e.,

$$TP_{SET} = \{t_i | (t_i, f_i) \in V_{TP}, U_1 \leq f_i \leq U_2\}, \quad (1)$$

where U_1 is a lower threshold obtained by a given neighbourhood percentage of TP (NTP), thus, $U_1 = (1 - NTP) * TP$. U_2 is the upper threshold and it is calculated in a similar way ($U_2 = (1 + NTP) * TP$). Both in WebCLEF 2005 and 2006, we have used $NTP = 0.4$, considering that the TP technique is language independent.

2.2 Term Enrichment

Certainly TP reduction may increase precision, but furthermore it decreases recall. Due to this fact, we enriched the selected terms by obtaining new terms, those with similar characteristics to the initial ones. Specifically, given a text T , with selected terms TP_{SET} , y is a new term if it co-occurs in T with some $x \in TP_{SET}$, i.e.,

$$TP'_{SET} = TP_{SET} \cup \{y | x \in TP_{SET} \wedge (fr(xy) > 1 \vee fr(yx) > 1)\}. \quad (2)$$

Considering the text length, we only selected a window of size 1 around each term of TP_{SET} , and a minimum frequency of two for each bigram was required as condition to include new terms.

2.3 Information Retrieval Model

Our information retrieval is based on the Boolean Model and, in order to rank the documents retrieved, we have used the Jaccard similarity function applied to both the query and to the documents within the collection. Previously, each document was preprocessed and its index terms were selected (the preprocessing phase is described in Section 3.1). As we will see in Section 3.3, we have represented each text by using the selection formula given in the Equation (1). Additionally, after the reduction step, we have carried out an enrichment process (see Equation (2)) based on the identification of related terms to those resulted from the selection.

3 Evaluation

3.1 Corpus

For the experiments carried out, we have used the EuroGOV corpus provided by the WebCLEF forum which is well described in [6]. We have indexed only 20 from 27 domains, namely: DE, AT, BE, DK, SI, ES, EE, IE, IT, SK, LU, MT, NL, LV, PT, FR, CY, GR, HU, and UK (we did not indexed the following domains: EU, RU, FI, PL, SE, CZ, LT). Due to this fact, only 1,470 from 1,939 topics were evaluated, which is approximately a 75.81% of the whole topics. Although we presented in Section 3.3 the MRR over 1,939 topics, there were 469 topics not indexed.

The preprocessing phase of the EuroGOV corpus was carried out by writing two scripts to obtain index terms for each document. The first script uses regular

expressions for excluding all the information which is enclosed by the characters < and >. Although this script obtains very good results, it is very slow and therefore we decided to use it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE).

On the other hand, we wrote a script based in the *html* syntax for obtaining all the terms considered interesting for indexing, i.e., those different than script codes (javascript, vbscript, style cascade sheet, etc), *html* codes, etc. This script speeded up our indexing process but it did not take into account that some web pages were incorrectly written and, therefore, we missed important information from those documents.

For every page compiled in the EuroGOV corpus, we also determine its language by using TexCat [8], a language identification program widely used. We constructed our evaluation corpus with those documents identified as a language of the above list.

Another preprocessing problem consisted in the charset codification, which led to an even more difficult analysis. Although the EuroGOV corpus is given in UTF-8, the documents that made up this corpus do not necessarily keep this charset. We have seen that for some domains, the charset codification is given in the *html* metadata tag, but also we found that very often this codification is wrong. We consider the charset codification detection the most difficult problem in the preprocessing step. Finally, we eliminated stopwords for each language (except for Greek language) and punctuation symbols. For the evaluation of this corpus, a set of queries was provided by WebCLEF 2006, which were applied with the same preprocessing process described above.

3.2 Indexing Reduction

After our first participation in WebCLEF [3], we carried out more experiments using only those documents in Spanish language from the EuroGOV corpus. We ob-

Table 1. Vocabulary size and percentage of reduction

Domain	DE	AT	BE	DK	SI
Size (KB)	2,588	2,317	6,796	1,189	6,729
Reduction (%)	95.3	97.2	98.0	97.9	97.1
Domain	ES	EE	IE	IT	SK
Size (KB)	16,271	4,838	2,632	11,913	14,668
Reduction (%)	98.5	97.2	96.0	98.4	97.5
Domain	LU	MT	NL	LV	PT
Size (KB)	3,212	4,817	20,324	21,213	9,134
Reduction (%)	99.2	95.7	97.7	97.8	97.6
Domain	FR	CY	GR	HU	UK
Size (KB)	22,083	18,814	340	10,440	14,239
Reduction (%)	95.8	96.5	97.4	98.8	96.1

served that a value of $NTP = 0.4$ using the reduction process shown in the Equation 1 was adequated. Therefore, in this test we carried out one run with that value. Moreover, this run took the evaluation corpus composed by the reduction of every text, using TP technique with a neighbourhood of 40% around TP, an enriched this set of terms using related terms as described by Equation (2).

Table 1 shows the size of each domain; the vocabulary composed by representation of all texts, $|TP'_{SET}|$, as well as the percentage of reduction obtained by each one with respect to the original vocabulary. As we can see, the TP technique obtained a vocabulary reduction percentage of more than 95%, which implies a time reduction for any search engine indexing process.

3.3 Results

Table 2 shows the results for the run submitted. The first and second columns indicate the number of topics evaluated and the test type. The last column shows the Mean Reciprocal Rank (MRR) obtained for each test. Additionally, the average success at different number of documents retrieved is shown; for instance, the third column indicates the average success of our search engine at the first answer.

The best overall results by using the *new topic set* is presented in Table 3. The results are reported on all topics (all), the automatic (auto) and manual (manual) subsets of topics. The average of the submitted run is shown in the fifth column.

Table 2. Evaluation results

#Topics	Test	Average Success at					Mean Reciprocal Rank
		1	5	10	20	50	
1939	All	0.0093	0.0217	0.0294	0.0371	0.0464	0.0157
1620	Auto	0.0025	0.0049	0.0086	0.0117	0.0160	0.0040
319	Man	0.0439	0.1066	0.1348	0.1661	0.2006	0.0750
810	A. bi.	0.0037	0.0062	0.0099	0.0123	0.0148	0.0049
124	M. new	0.0323	0.0968	0.1129	0.1613	0.2339	0.0657
810	A. uni.	0.0012	0.0037	0.0074	0.0111	0.0173	0.0031
195	M. old	0.0513	0.1128	0.1487	0.1692	0.1795	0.0810

Table 3. Results by using the new topic set

Run	all	auto	manual	average
allpt40bi	0.0272	0.0080	0.0790	0.0435

4 Conclusions

We have proposed an index reduction method for cross language search engines, which includes an enrichment step. Our proposal is based on the transition point technique which allows to index only the mid-frequency terms from every

document. Our method is linear in computational time and, therefore, it can be used in a wide spectrum of practical tasks.

After submitting our run we observed enhancement with respect to the results obtained in the BiEnEs task in WebCLEF 2005 [3]. WebCLEF 2005. By using the enrichment, more than 40% improvement on MRR was achieved. However, by using the Vector Space Model similar results to the boolean model were obtained.

The TP technique has shown an effective use on diverse areas of NLP, and its best features for NLP, are mainly two: a high content of semantic information and the sparseness that can be obtained on vectors for document representation on models based on the vector space model. On the other hand, its language independence allows to use this technique in multilingual environments.

We consider that our approach may be improved by taking into account all the terms of the vocabulary in the enrichment process. Once the term expansion would be done, the mid-frequency selection technique could be applied. Further analysis will investigate this issue.

References

1. Booth, A.: A Law of Occurrences for Words of Low Frequency. Information and control (1967)
2. Balog, K., Azzopardi, L., Kamps, J., de Rijke, M.: Overview of WebCLEF 2006. In: This Volume (2007)
3. Pinto, D., Jiménez-Salazar, H., Rosso, P., Sanchis, E.: BUAP-UPV TPIRS: A System for Document Indexing Reduction at WebCLEF, in Accessing Multilingual Information Repositories. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. CLEF 2005: Cross-Language Evaluation Forum (2005), <http://www.clef-campaign.org/>
5. Reyes-Aguirre, B., Moyotl-Hernández, E., Jiménez-Salazar, H.: Reducción de Términos Índice Usando el Punto de Transición. In: proceedings of Facultad de Ciencias de Computación XX Anniversary Conferences, BUAP (2003)
6. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: EuroGOV: Engineering a Multilingual Web Corpus. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
7. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: WebCLEF 2005: Cross-Lingual Web Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. TextCat: Language identification tool (2005), <http://odur.let.rug.nl/~vannord/TextCat/>
9. Urbizagástegui, R.: Las posibilidades de la Ley de Zipf en la indización automática, Research report of the California Riverside University (1999)
10. Zipf, G.K.: Human Behavior and the Principle of Least-Effort. Addison-Wesley, Cambridge, MA (1949)

Experiments with the 4 Query Sets of WebCLEF 2006

Stephen Tomlinson

Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
<http://www.opentext.com/>

Abstract. In the WebCLEF mixed monolingual retrieval task of the Cross-Language Evaluation Forum (CLEF) 2006, the system was given 1939 known-item queries, and the goal was to find the desired page in the 82GB EuroGOV collection (3.4 million pages crawled from government sites of 27 European domains). The 1939 queries included 124 new manually-created queries, 195 manually-created queries from last year, and two sets of 810 automatically-generated queries. In our experiments, the results on the automatically-generated queries were not always predictive of the results on the manually-created queries; in particular, our title-weighting and duplicate-filtering techniques were fairly effective on the manually-created queries but were detrimental on the automatically-generated queries. Further investigation uncovered serious encoding issues with the automatically-generated queries; for instance, the queries labelled as Greek actually used Latin characters.

1 Introduction

The WebCLEF 2006 mixed monolingual task used 4 query sets, herein denoted as “new” (124 new manually-created queries), “old” (195 manually-created queries from last year), “uni” (810 automatically-generated queries via a unigram model) and “bi” (810 automatically-generated queries via a bigram model). Details of the topic (query) creation are in [1].

For each query set, we performed 4 runs, herein denoted as “base” (plain content search), “p” (same as “base” except that it put additional weight on matches in the title, url, first heading and some meta tags, including extra weight on matching the query as a phrase in these fields), “dp” (same as “p” except that it put additional weight on urls of depth 4 or less), and “dpc” (same as “dp” except that an experimental duplicate-filtering heuristic was applied). The details of indexing and ranking were similar to what we described last year [4].

The focus of this (short) paper is to investigate whether the results on the experimental automatically-generated queries successfully predict the results on the (presumably realistic) manually-created queries.

2 Results

Table 1 lists the mean scores of the 4 runs for each query set. The measures listed are as follows:

- *Generalized Success@10* (GS10): For a topic, GS10 is 1.08^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found. (Last year [4], GS10 was known as “First Relevant Score” (FRS).)
- *Success@n* (S@n): For a topic, Success@n is 1 if a desired page is found in the first n rows, 0 otherwise. The table includes Success@1 (S1), Success@5 (S5), Success@10 (S10) and Success@50 (S50).
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found. “Mean Reciprocal Rank” (MRR) is the mean of the reciprocal ranks over all the topics.

We focus on 2 web techniques which are denoted as follows:

- ‘P’ (extra weight for the Title, etc.): The “p” score minus the “base” score.
- ‘C’ (duplicate-filtering): The “dpc” score minus the “dp” score.

Table 1. Mean Scores of Submitted WebCLEF Runs

Run	GS10	S1	S5	S10	S50	MRR
new-dpc	0.657	49/124	75/124	84/124	102/124	0.494
new-dp	0.665	49/124	78/124	86/124	105/124	0.497
new-p	0.666	49/124	78/124	86/124	104/124	0.499
new-base	0.648	44/124	73/124	84/124	101/124	0.466
old-dpc	0.610	71/195	110/195	122/195	154/195	0.455
old-dp	0.600	71/195	107/195	119/195	159/195	0.447
old-p	0.571	67/195	101/195	115/195	154/195	0.425
old-base	0.528	59/195	94/195	104/195	143/195	0.390
uni-dpc	0.089	21/810	59/810	72/810	116/810	0.048
uni-dp	0.115	21/810	69/810	92/810	177/810	0.056
uni-p	0.113	23/810	65/810	93/810	172/810	0.056
uni-base	0.116	24/810	64/810	92/810	183/810	0.057
bi-dpc	0.078	24/810	51/810	65/810	107/810	0.046
bi-dp	0.091	24/810	55/810	73/810	148/810	0.050
bi-p	0.088	23/810	49/810	70/810	148/810	0.048
bi-base	0.091	23/810	56/810	70/810	148/810	0.049

Table 2 shows, for each of the 4 query sets, the impact of the 2 web techniques on Generalized Success@10 (GS10). Its columns are as follows:

Table 2. Impact of Web Techniques on Generalized Success@10 (GS10)

Expt	Δ GS10	95% Conf	vs.	3 Extreme Diffs (Topic)
new-P	0.018	(−0.015, 0.051)	26-24-74	0.73 (1785), 0.63 (1420), −0.54 (630)
new-C	−0.009	(−0.035, 0.017)	20-4-100	−0.86 (1310), −0.79 (1163), 0.26 (831)
old-P	0.043	(0.017, 0.070)	61-28-106	1.00 (891), 0.86 (729), −0.56 (922)
old-C	0.010	(−0.011, 0.031)	43-7-145	−0.86 (865), −0.74 (516), 0.57 (794)
uni-P	−0.003	(−0.007, 0.002)	36-80-694	0.63 (1023), −0.41 (1410), −0.63 (1833)
uni-C	−0.025	(−0.036, −0.015)	42-66-702	−0.93 (1417), −0.93 (235), 0.54 (1833)
bi-P	−0.002	(−0.007, 0.002)	24-58-728	1.00 (154), −0.63 (569), −0.64 (1580)
bi-C	−0.013	(−0.021, −0.005)	45-50-715	−0.93 (548), −0.93 (1005), 0.70 (1076)

- “Expt” specifies the experiment (query set and web technique using the codes defined earlier).
- “ Δ GS10” is the increase in the mean GS10 score from the web technique.
- “95% Conf” is an approximate 95% confidence interval for Δ GS10 (calculated from plus/minus twice the standard error of the mean difference). If zero is not in the interval, the result is “statistically significant” (at the 5% level), i.e. the feature is unlikely to be of neutral impact (on average), though if the average difference is small (e.g. <0.020) it may still be too minor to be considered “significant” in the magnitude sense.
- “vs.” is the number of queries on which the technique increased the score, decreased the score and produced the same score (respectively). These numbers should always add to the number of queries.
- “3 Extreme Diffs (Topic)” lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

The “P” and “C” techniques are useful techniques for web retrieval in our experience, and Table 2 shows that they are beneficial on the manually-created query sets (“new” and “old”). (Although the mean score for “C” was down on the “new” queries, 20 queries were improved compared to just 4 declines, and we suspect some of the large declines may be from the new official judgements failing to mark all of the duplicates, though we have not carefully checked at this time.)

The automatically-generated query sets (“uni” and “bi”) suggest that the “P” and “C” techniques are detrimental, which is evidence that the automatically-generated queries can produce misleading results for web retrieval.

2.1 Encoding Issues

A challenge of working with the WebCLEF document set (and hence for automatically producing queries) is determining how to convert the document text to Unicode. When we looked at the automatically-generated Greek queries, we found that they still used the Latin alphabet. For example, topic WC0600153 is labelled as domain="gr" and language="EL", but the query is "ufp resjiyusjp uxnjdp". Last year's manually-created Greek queries were in proper Greek; e.g. the query for last year's Greek topic WC0112 was Πλήρης λίστα των υπουργών και υφυπουργών όλων υπουργείων της Ελληνικής κυβέρνησης (List of ministers and deputy ministers for all the ministries of the Greek government).

We suspect there could be several topics with encoding problems in the automatically-generated topic set, not just for Greek, but also for languages which use Cyrillic (such as Russian) and perhaps for Eastern European languages. Even in Latin-based languages, any topic with non-ASCII characters (such as accented characters) should probably be reviewed.

3 Conclusions

We have demonstrated that the automatically-generated queries of WebCLEF 2006 can produce misleading results for web retrieval. In particular, our title-weighting and duplicate-filtering techniques, which we believe are effective techniques and which were effective on the manually-created queries, were detrimental on the automatically-generated queries. Further investigation uncovered serious encoding issues with the automatically-generated queries; for instance, the queries labelled as Greek actually used Latin characters this year.

We are skeptical that useful query sets can be created automatically for research tasks. A lot of the research challenges of web retrieval, such as proper title weighting, duplicate detection and character encoding detection, need to be resolved before query sets can be created automatically. And if all of those challenges were resolved, there would no longer be a research need for producing the query set (it would just be of routine testing value). At present, perhaps the best that could be done would be to automatically enhance the assessments of the manual topics (from last year and this year) to include all duplicates (something that reportedly has been done for known-item query sets in the TREC Terabyte task [2]).

References

1. Balog, K., Azzopardi, L., Kamps, J., de Rijke, M.: Overview of WebCLEF 2006. (In this volume) (2007)
2. Clarke, C.L.A., Scholer, F., Soboroff, I.: The TREC 2005 Terabyte Track. In: Proceedings of TREC 2005 (2006)
3. Tomlinson, S.: Comparing the Robustness of Expansion Techniques and Retrieval Measures. (In this volume) (2007)
4. Tomlinson, S.: European Web Retrieval Experiments with Hummingbird SearchServerTM at CLEF 2005. Working Notes of CLEF (2005)

Applying Relevance Feedback for Retrieving Web-Page Retrieval

Syntia Wijaya, Bimo Widhi, Tommy Khoerniawan, and Mirna Adriani

Faculty of Computer Science

University of Indonesia

Depok 16424, Indonesia

{swd20, bimo20, tokh20}@mhs.cs.ui.ac.id, mirna@cs.ui.ac.id

Abstract. We present a report on our participation in the mixed monolingual web task of the 2006 Cross-Language Evaluation Forum (CLEF). We compared the result of web page retrieval based on the page content, page title, and anchor page. The retrieval effectiveness for the combination of page content, page title, and anchor texts was better than that of the combination of page title and page title only. Applying the pseudo-relevance feedback improved the retrieval performance of the queries.

Keywords: web retrieval, relevance feedback.

1 Introduction

The fast growing amount of information on the web motivated many researchers to come up with a way to deal with such information efficiently [2, 5]. Information retrieval forums such as the Cross Language Evaluation Forum (CLEF) have included research in the web area. In fact, since 2005, CLEF includes a WEBIR topic as one of the research tracks. This year we, the University of Indonesia IR-Group, participated in the mixed monolingual WEBIR - CLEF 2006 task.

The web page contains some characteristics that the newspaper document does not have such as the structure of the web page, anchor text, or URL length. These characteristics have been studied to have positive effects on the retrieval of web pages [2]. This paper reports our attempt to study whether these characteristics can result in better retrieval performance instead of using word frequency alone. We also study if applying a word stemmer to web documents can increase the retrieval performance as with newspaper documents [1].

2 The Retrieval Process

The mixed monolingual task searches for web pages in a number of languages. The queries and the documents were processed using the *Lemur*¹ information retrieval system. Stop-word removal, as is done by many IR systems, was applied only to the

¹ See “<http://www.lemurproject.org>”.

English queries and documents. However the stemmer is not used in this experiment since our previous work shows that it did not help in improving the retrieval result.

2.1 Web Page Scoring Techniques

We employed five different techniques for scoring the relevance of documents (web pages) in the collection, i.e., based on the combination of the content of the page, the title of the page, and the anchor texts that appear on the pages.

The first technique takes into account only the content of a web page to find the most relevant web pages to the query. We used the *language model* [3] to find the probability value between the query and the pages. The second technique considers the title of the web page as the only source in finding the relevant pages. The third technique uses the content and the title of the page to find the relevant pages.

3 Experiment

The web collection contains over two million documents from the EUROGOV collection. In these experiments, we used *Lemur* information retrieval system to index and retrieve the documents. *Lemur* is built based on the *language model* [3]. We index the webpages according to their content pages, title pages, and anchors. Stopwords were removed from the collection, but word stemming was not applied to the collection.

In this experiment we expanded the query using the pseudo relevance feedback [4] which is known as a technique to improve the retrieval performance. We used BM25 weighting formula [1] to find 5 terms from the top-5 documents.

4 Results

The first result is shown in Table 1. In the retrieval, we compute the total relevance score by summing up the relevance scores based on page content, page title, and anchor texts found on the web-pages.

Table 1. Mean Reciprocal Mean (MRR) of the combined relevance score for page content, page title, and anchor texts on a webpage

Task : Mixed Monolingual	UI1DTA
MRR	0.0404
Average success at 1:	0.0258
Average success at 5:	0.0531
Average success at 10:	0.0707

Table 2 shows the result of combining the relevance scores based on page content and page title. As can be seen, the MRR dropped 71.28% from 0.0404 (see Table 1) to 0.0116 (see Tabel 2).

Table 2. Mean Reciprocal Mean (MRR) of the combined relevance score for page content and page title

Task : Mixed Monolingual	UI4DTW
MRR	0.0116
Average success at 1:	0.0067
Average success at 5:	0.0150
Average success at 10:	0.0201

The third technique applies the pseudo-relevance feedback to the retrieval that uses the combined score of page content, page title, and anchor texts. As shown in Table 3, the feedback reduced the performance of the queries where the MRR dropped to 0.0253. The pseudo-relevance feedback was done using the top-5 relevant documents retrieved.

Table 3. Mean Reciprocal Mean (MRR) of the combined score of page content, page title, and anchor texts with top-5 documents pseudo-relevance feedback

Task : Mixed Monolingual	UI3DTAF
MRR	0.0253
Average success at 1:	0.0160
Average success at 5:	0.0309
Average success at 10:	0.0423

Finally, the last result was obtained by applying the pseudo-relevance feedback to the combined relevance score of page content and page title only. As shown in Table 4, we obtained the highest retrieval performance with MRR of 0.0918. The top-5 documents were used to find 5 terms that were added to the query.

To investigate the cause of our poor retrieval performance, we conducted some further experiments. We used the queries from last year’s task and ran them on the same index that was built using *Lemur*. The result is as shown in Table 5, which is much better than for this year’s queries. However, we found a sign of indexing error,

Table 4. Mean Reciprocal Mean (MRR) of the combined score of page content and page title with top-5 documents pseudo-relevance feedback

Task : Mixed Monolingual	UI1DTF
MRR	0.0918
Average success at 1:	0.0634
Average success at 5:	0.1202
Average success at 10:	0.1516

Table 5. Mean Reciprocal Mean (MRR) of the combined relevance score of page content, page title, and anchor texts using the 2005 query-topics

Task : Mixed Monolingual	DTA- 2005
MRR	0.2069
Average success at 1:	0.1444
Average success at 5:	0.2742
Average success at 10:	0.3254

i.e., there were some domains that *Lemur* was unable to index. This resulted in Lemur's not being able to retrieve any documents for a number of queries. We also suspected that the index for documents in languages containing non-latin characters was corrupt, as indicated by the fact that documents in some domains such as Russian and Greek were never retrieved.

5 Summary

Our results demonstrate that combining the page content, the page title, and anchor texts resulted in a better mean reciprocal rank (MRR) compared to searching using the page content and page title only. The pseudo-relevance feedback that we employed increased the retrieval performance of the queries. We hope to improve our results in the future by exploring still other methods.

References

1. Richardo, B.-Y., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, New York (1999)
2. Hawking, D.: Overview of the TREC-9 Web Track. In: NIST Special Publication: The 10th Text Retrieval Conference (TREC-10) (2001)
3. Ponte, J., Croft, W.B.A.: Language Modeling Approach to Information Retrieval. In: Proceedings of the 21st ACM SIGIR Conference on Research and development in Information Retrieval, pp. 275–281. ACM Press, New York (1998)
4. Salton, Gerard, McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
5. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of ACM SIGIR'98, Melbourne, Australia, ACM, New York (1998)

GeoCLEF 2006: The CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview

Fredric Gey¹, Ray Larson¹, Mark Sanderson², Kerstin Bischoff³, Thomas Mandl³,
Christa Womser-Hacker³, Diana Santos⁴, Paulo Rocha⁴, Giorgio M. Di Nunzio⁶,
and Nicola Ferro⁶

¹ University of California, Berkeley, CA, USA
gey@berkeley.edu, ray@sims.berkeley.edu

² Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@sheffield.ac.uk

³ Information Science, University of Hildesheim, Germany
{mandl, womser}@uni-hildesheim.de

⁴ Linguateca, SINTEF ICT, Norway
Diana.Santos@sintef.no, Paulo.Rocha@di.uminho.pt

⁶ Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it

Abstract. After being a pilot track in 2005, GeoCLEF advanced to be a regular track within CLEF 2006. The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR): retrieval for topics with a geographic specification. For GeoCLEF 2006, twenty-five search topics were defined by the organizing groups for searching English, German, Portuguese and Spanish document collections. Topics were translated into English, German, Portuguese, Spanish and Japanese. Several topics in 2006 were significantly more geographically challenging than in 2005. Seventeen groups submitted 149 runs (up from eleven groups and 117 runs in GeoCLEF 2005). The groups used a variety of approaches, including geographic bounding boxes, named entity extraction and external knowledge bases (geographic thesauri and ontologies and gazetteers).

1 Introduction

Existing evaluation campaigns such as TREC and CLEF have not, prior to 2005, explicitly evaluated geographical relevance. The aim of GeoCLEF is to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects. Participants are offered a TREC style ad hoc retrieval task based on existing CLEF collections. GeoCLEF 2005 was run as a pilot track to evaluate retrieval of multilingual documents with an emphasis on geographic search on English and German document collections. Results were promising, but it was felt that more work needed to be done to identify the research and evaluation issues surrounding geographic information retrieval from text. Thus 2006 was the second year in which GeoCLEF was run as a track within CLEF. For 2006, two additional document languages were added to GeoCLEF, Portuguese and Spanish. GeoCLEF was a

collaborative effort by research groups at the University of California, Berkeley (USA), the University of Sheffield (UK), University of Hildesheim (Germany), Linguateca (Norway and Portugal), and University of Alicante (Spain). Seventeen research groups (increased from eleven in 2005) from a variety of backgrounds and nationalities submitted 149 runs (up from 117 in 2005) to GeoCLEF.

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Given that many documents contain some kind of spatial reference, there are examples where geographical references (geo-references) may be important for IR. For example, to retrieve, re-rank and visualize search results based on a spatial dimension (e.g. “find me news stories about riots near Dublin City”). In addition to this, many documents contain geo-references expressed in multiple languages which may or may not be the same as the query language. For example, the city of Cologne (English) is also Köln (German), Colónia in Portuguese from Portugal, Colônia in Brazilian Portuguese, and Colonia (Spanish). Queries with names such as this may require an additional translation step to enable successful retrieval.

For 2006, Spanish and Portuguese, in addition to German and English, were added as document languages, while topics were developed in all four languages with topic translations provided for the other languages. In addition the National Institute of Informatics of Tokyo, Japan translated the English version of the topics to Japanese. There were two Geographic Information Retrieval tasks: monolingual (English to English, German to German, Portuguese to Portuguese and Spanish to Spanish) and bilingual (language X to language Y, where X or Y was one of English, German, Portuguese or Spanish and additionally X could be Japanese).

2 Document Collections Used in GeoCLEF

The document collections for this year's GeoCLEF experiments are all newswire stories from the years 1994 and 1995 used in previous CLEF competitions. Both the English and German collections contain stories covering international and national news events, therefore representing a wide variety of geographical regions and places. The English document collection consists of 169,477 documents and was composed of stories from the British newspaper *The Glasgow Herald* (1995) and the American newspaper *The Los Angeles Times* (1994). The German document collection consists of 294,809 documents from the German news magazine *Der Spiegel* (1994/95), the German newspaper *Frankfurter Rundschau* (1994) and the Swiss news agency SDA (1994/95). Although there are more documents in the German collection, the average document length (in terms of words in the actual text) is much larger for the English collection. In both collections, the documents have a common structure: newspaper-specific information like date, page, issue, special filing numbers and usually one or more titles, a byline and the actual text. The document collections were not geographically tagged or contained any other location-specific information. For Portuguese, GeoCLEF 2006 utilized two newspaper collections, spanning over 1994-1995, for respectively the Portuguese and Brazilian newspapers *Público* (106,821 documents) and *Folha de São Paulo* (103,913 documents). Both are major daily newspapers in their countries. Not all material published by the two newspapers is included

in the collections (mainly for copyright reasons), but every day is represented. The collections are also distributed for IR and NLP research by Linguatca as the CHAVE collection (www.linguatca.pt/CHAVE/, see URL for DTD and document examples). The Spanish collection was composed of Spanish newspapers EFE 1994-1995 distributed by the Spanish Agency EFE (<http://www.efc.es/>). EFE 1994 are made up of 215,738 documents and EFE 1995 of 238,307 documents.

3 Generating Search Topics

A total of 25 topics were generated for this year's GeoCLEF. Topic creation was shared among the four organizing groups, each group creating initial versions of their proposed topics in their language, with subsequent translation into English. In order to support topic development, Ray Larson indexed all collections with his Cheshire II document management system and this was made available to all organizing groups for interactive exploration of potential topics. While the aim had been to prepare an equal number of topics in each language, ultimately only two topics (GC026 and GC027) were developed in English. Other original language numbers were German, 8 topics (GC028 to GC035), Spanish, 5 topics (GC036 to GC040) and Portuguese, 10 topics (GC041 to GC050). This section will discuss the creation of the spatially-aware topics for the track.

3.1 Topic Generation

In GeoCLEF 2005 some criticism arose about the lack of geographical challenges of the topics (favouring keyword-based approaches) and the German task was inherently more difficult because several topics had no relevant documents in the German collections. Therefore geographical and cross-lingual challenge and equal distribution across language collections was considered central during topic generation. Topics should vary according to the granularity and kind of geographic entity and should require adequate handling of named entities within the process of translation (e.g. regarding decompounding, transliteration or translation).

For English topic generation, Fred Gey simply took two topics he had considered in the past (Wine regions around rivers in Europe and Cities within 100 kilometers of Frankfurt, Germany) and developed them. The latter topic (GC027) evolved into an exact specification of the latitude and longitude of Frankfurt am Main (to distinguish it from Frankfurt an der Oder) in the narrative section. Interactive exploration verified that documents could be found which satisfied these criteria on the basis of geographic knowledge by the proposer (i.e. the Rhine and Moselle valleys of Germany and cities Heidelberg, Koblenz, Mainz, and Mannheim near Frankfurt).

The German group at Hildesheim started with brain storming on interesting geographical notions and looking for potential events via the Cheshire II Interface, we unfortunately had to abandon all smaller geographic regions soon. Even if a suitable number of relevant documents could be found in one collection, most times there were few or no respective documents in the other language collections. This may not be surprising, because within the domain of news criteria like (inter)national relevance, prominence, elite nation or elite person besides proximity, conflict/negativism

and continuity etc. (for an overview see Eidlers[2]) are assumed to affect what will become a news article. Thus, the snow conditions or danger of avalanches in Grisons (canton in Switzerland) may be reported frequently by the Swiss news agency *SDA* or even German newspapers, whereas the British or American newspapers may not see the relevance for their audience. In addition, the geographically interesting issue of tourism in general is not well represented in the German collection. As a result well known places and larger regions as well as international relevant or dramatic concepts had to be focused on, although this may not reflect all user needs for GIR systems (see also Kluck & Womser-Hacker[7]). In order not to favor systems relying purely on keywords we concentrated on more difficult geographic entities like historical or political names used to refer geographically to a certain region and imprecise regions like *the Ruhr* or *the Middle East*. Moreover some topics should require the use of external geographic knowledge e.g. to identify cities onshore of the *Sea of Japan* or *the Tropics*. The former examples introduce ambiguity or translation challenges as well. *Ruhr* could be the river or the area in Germany and *the Middle East* may be translated to German *Mittlerer Osten*, which is nowadays often used, but would denote a slightly different region. The naming of the *Sea of Japan* is difficult as it depends on the Japanese and Western perspective, whereas in Korea it would be named *East Sea (of Korea)*. After checking such topic candidates for relevant documents in other collections we proposed eight topics, which we thought would contribute to a topic set varying in thematic content and system requirements.

The GeoCLEF topics proposed by the Portuguese group (a total of 10) were discussed between Paulo Rocha and Diana Santos, according to an initial typology of possible geographical topics (see below for a refined one) and after having scrutinized the frequency list of proper names in both collections, manually identifying possible places of interest. Candidate topics were then checked in the collections, using the Web interface to the AC/DC project [8], to investigate whether they were well represented. We included some interesting topics from a Portuguese (language) standpoint, including "ill-defined" or at least little known regions in an international context, such as *norte de Portugal* (North of Portugal) or *Nordeste brasileiro* (Brazilian Northeast). Basically, they are very familiar and frequently used concepts in Portuguese, but have not a purely geographical explanation. Rather, they have a strongly cultural and historical motivation. We also inserted a temporally dependent topic (outdated Champion's Cup, now Champion's League – and already in 1994-1995 as well, but names continue their independent life in newspapers and in folk's stock of words). This topic is particularly interesting, since it in addition concerns "European" football, where one of the partners is (non-geographically-European) Israel.

We also strove to find topics which included more geographical relations than mere "in" (homogeneous region), as well as different location types as far as grain and topology are concerned. As to the first concern, note that although "shipwrecks in the Atlantic Ocean" seem to display an ordinary "in"-relation, shipwrecks are often near the coasts, and the same is still more applicable about topics such as "fishing in Newfoundland", where it is presupposed that you fish on the sea near the place (or that you are concerned with the impact of fishing to Newfoundland). Likewise, anyone who knows what ETA stands for would at once expect that "ETA's activities in France" would be mostly located in the French Basque country (and not anywhere in France).

For the second concern, that of providing different granularity and/or topology, note that the geographical span of forest fires is clearly different from that of lunar or solar eclipses (a topic suggested by the German team). As to form of the region, the "New England universities" topic circumscribes the "geographical region" to a set of smaller conceptual "regions", each represented by a university. Incidentally, this topic displays another complication, because it involves a multiword named entity: not only "New England" is made up of two different words but both are very common and have a specific meaning on its own (in English and Portuguese alike). This case is further interesting because it would be as natural to say *New England* in Portuguese as *Nova Inglaterra*, given that the name is not originally Portuguese.

We should also report interesting problems caused by translation into Portuguese from topics originally stated in other languages (they are not necessarily translation problems, but were spotted because we had to look into the particular cases of those places or expressions). For example, Middle East can be equally translated by *Próximo Oriente* and *Médio Oriente*, and it is not politically neutral how precisely in that area some places are described. For example, we chose to use the word *Palestina* (together with *Israel*) and leave out Gaza Strip (which, depending on political views, might be considered a part of both). What is interesting here is that the political details are absolutely irrelevant for the topic in question (which deals with archaeological findings), but the GeoCLEF organizers (in common) decided to specify a lower level, or a higher precision description, of every location/area mentioned, in the narrative, so that a list of Middle East countries and regions had to be supplied, and agreed upon.

3.2 Format of Topic Description

The format of GeoCLEF 2006 differed from that of 2005. No explicit geographic structure was used this time, although such a structure was discussed by the organizing groups. Two example topics are shown in Figure 1.

```
<top>
<num>GC027</num>
<EN-title>Cities within 100km of Frankfurt</EN-title>
<EN-desc>Documents about cities within 100 kilometers of the city of Frankfurt in
Western Germany</EN-desc>
<EN-narr>Relevant documents discuss cities within 100 kilometers of Frankfurt am Main
Germany, latitude 50.11222, longitude 8.68194. To be relevant the document must describe
the city or an event in that city. Stories about Frankfurt itself are not relevant</EN-
narr>
</top>
<top>
<num> GC034 </num>
<EN-title> Malaria in the tropics </EN-title>
<EN-desc> Malaria outbreaks in tropical regions and preventive vaccination </EN-desc>
<EN-narr> Relevant documents state cases of malaria in tropical regions and possible
preventive measures like chances to vaccinate against the disease. Outbreaks must be of
epidemic scope. Tropics are defined as the region between the Tropic of Capricorn,
latitude 23.5 degrees South and the Tropic of Cancer, latitude 23.5 degrees North. Not
relevant are documents about a single person's infection.</EN-narr>
</top>
```

Fig. 1. Topics GC027: Cities within 100 Kilometers of Frankfurt and GC034: Malaria in the Tropics

As can be seen, after the brief descriptions within the title and description tags, the narrative tag contains detailed description of the geographic detail sought and the relevance criteria.

3.3 Several Kinds of Geographical Topics

We came up with a tentative classification of topics according to the way they depend on place (in other words, according to the way they can be considered "geographic"), which we believe to be one of the most interesting results of our participation in the choice and topic formulation for GeoCLEF. Basically, this classification was done as an answer to the overall too simplistic assumption of first GeoCLEF[3], namely the separation between subject and location as if the two were independent and therefore separable pieces of information. (Other comments to the unsuitability of the format used in GeoCLEF can be found in Santos and Cardoso [9], and will not be repeated here.)

While it is obvious that in some (simple) cases geographical topics can be modeled that way, there's much more to place and to the place of place in the meaning of a topic than just that, as we hope this categorization can help making clear:

- 1 non-geographic subject restricted to a place (music festivals in Germany) [only kind of topic in GeoCLEF 2005]
- 2 geographic subject with non-geographic restriction (rivers with vineyards) [new kind of topic added in GeoCLEF 2006]
- 3 geographic subject restricted to a place (cities in Germany)
- 4 non-geographic subject associated to a place (independence, concern, economic handlings to favour/harm that region, etc.) Examples: *independence of Quebec*, *love for Peru* (as often remarked, this is frequently, but not necessarily, associated to the metonymical use of place names)
- 5 non-geographic subject that is a complex function of place (for example, place is a function of topic) (*European football cup matches*, *winners of Eurovision Song Contest*)
- 6 geographical relations among places (*how are the Himalayas related to Nepal?* Are they inside? Do the Himalaya mountains cross Nepal's borders? etc.)
- 7 geographical relations among (places associated to) events (*Did Waterloo occur more north than the battle of X?* *Were the findings of Lucy more to the south than those of the Cromagnon in Spain?*)
- 8 relations between events which require their precise localization (*was it the same river that flooded last year and in which killings occurred in the XVth century?*)

Note that we here are not even dealing with the obviously equally relevant interdependence of the temporal dimension, already mentioned above, and which was actually extremely conspicuous in the preliminary discussions among this year's organizing teams, concerning the denotation of "former Eastern bloc countries" and "former Yugoslavia" now (that is, in 1994-1995). In a way, as argued in Santos and Chaves[10], which countries or regions to accept as relevant depends ultimately on the user intention (and need). Therefore, pinning down the meaning of a topic depends on geographical, temporal, cultural, and even personal constraints, that are intertwined in

a complex way, and more often than not do not allow a clear separation. To be able to make sense of these complicated interactions and arrive at something relevant for a user by employing geographical reasoning seems one of the challenges that lies ahead in future GeoCLEF tracks.

4 Approaches to Geographic Information Retrieval

The participants used a wide variety of approaches to the GeoCLEF tasks, ranging from basic IR approaches (with no attempts at spatial or geographic reasoning or indexing) to deep NLP processing to extract place and topological clues from the texts and queries. Specific techniques used included:

- Ad-hoc techniques (blind feedback, German word compounding, manual query expansion)
- Gazetteer construction (GNIS, World Gazetteer)
- Gazetteer-based query expansion
- Question-answering modules utilizing passage retrieval
- Geographic Named Entity Extraction
- Term expansion using Wordnet
- Use of geographic thesauri (both manually and automatically constructed)
- Resolution of geographic ambiguity
- NLP – part-of-speech tagging

5 Relevance Assessment

English assessment was shared by Berkeley and Sheffield Universities. German assessment was done by the University of Hildesheim, Portuguese assessment by Linguateca, and Spanish assessment by University of Alicante. All organizing groups utilized the DIRECT System provided by the University of Padua. The Padua system allowed for automatic submission of runs by participating groups and for automatic assembling of the GeoCLEF assessment pools by language.

5.1 English Relevance Assessment

The English document pool extracted from 73 monolingual and 12 bilingual (language X to) English runs consisted of 17,964 documents to be reviewed and judged by our 5 assessors or about 3,600 documents per assessor. In order to judge topic GC027 (Cities within 100km of Frankfurt), Ray Larson used data from the GeoNames Information System along with the Cheshire II geographic distance calculation function, to extract and prepare a spreadsheet of populated places whose latitude and longitude was within a distance of 100 km of the latitude and longitude of Frankfurt. This spreadsheet contained 5342 names and was made available to all groups doing assessment. If a document in the pool contained the name of a German city or town, it was checked against the spreadsheet to see if it was within 100km of Frankfurt. Thus documents with well-known names (Mannheim, Heidelberg) were easily recognized, but Mecklenberg (where the German Grand Prix auto race is held) was not so easily

recognized. In reading the documents in the pool, we were surprised to find many Los Angeles Times documents about secondary school sports events and scores in the pool. A closer examination revealed that these documents contained the references to American students who had the same *family name* as German cities and towns. It is clear that geographic named entity disambiguation from text still needs some improvement.

5.2 German Relevance Assessment

For the pool of German monolingual and bilingual runs X2German 14,094 documents from the newspaper *Frankfurter Rundschau*, the Swiss news agency *SDA* and the news magazine *Spiegel* had to be assessed. Every assessor had to judge a number of assigned topics. Decisions on dubious cases were left open and then discussed within the group and/or the other language co-ordinators. Since many topics had clear, predefined criteria as specified in title, description and narrative, searching first the key concepts and their synonyms within the documents and then identifying their geographical reference led to rejecting the bulk of documents as irrelevant. Depending on the geographic entity asked for, manual expansion, e.g., the country names of the Middle East and their capitals, was done to query the DIRECT System provided by the University of Padua. Of course, such a list could never be complete and available resources would not be comprehensive enough to capture all possible expansions (e.g. we could not verify the river Code on the island of Java). Thus skimming over the text was often necessary to capture the documents main topic and geographical scope.

While judging relevance was generally easier for the short news agency articles of *SDA* with their headlines, keywords and restriction to one issue, *Spiegel* articles took rather long to judge, because of their length and essay-like stories often covering multiple events etc. without a specific narrow focus. Many borderline cases for relevance resulted from uncertainties about how broad/narrow a concept term should be interpreted and how explicit the concept must be stated in the document (e.g. do parked cars destroyed by a bomb correspond to a car bombing? Are attacks on foreign journalists and the Turkish invasion air attacks to be considered relevant as fulfilling the concept of combat?). Often it seems that for a recurring news issue it is assumed that facts are already known, so they are not explicitly cited. To keep the influence of order effects minimal is critical here.

Similarly, assessing relevance regarding the geographical criterion brought up a discussion on specificity *wrt* implicit inclusion. In all cases, reference to the required geographic entity had to be explicitly made, i.e., a document reporting about Fishing in the Norwegian Sea or the Greenland Sea without mentioning e.g. a certain coastal city in Greenland or Newfoundland was not considered relevant. Moreover, the borders of oceans and its minor seas are often hard to define (e.g. does Havana, Cuba border the Atlantic Ocean?). Figuring out the location referred to was frequently difficult, when the city mentioned first in an article could have been the domicile of the news agency or/and the city some event occurred in. This was especially true for GC040 *active volcanoes* and for GC027 *cities within 100km from Frankfurt*, with Frankfurt being the domicile of the *Frankfurter Rundschau*, which formed part of the collection. Problems with fuzzy spatial relations or imprecise regions on the other hand did not figure very prominently as they were defined in the extended narratives

(e.g. “near” Madrid includes only Madrid and its outskirts) and the documents to be judged did not contain critical cases. However, one may have argued against the decision to exclude all districts of Frankfurt as they do not form own cities, but have a common administration.

The topic on *cities around Frankfurt* (GC027) was together with GC050 about *cities along the Danube and the Rhine* the most difficult one to judge. Although a list of relevant cities containing more than 4000 names was provided by Ray Larson, this could not be used efficiently for relevance assessment to query the DIRECT system. Moreover, the notion of an event or a description made assessment even more time-consuming. We queried about 40 or 50 prominent relevant cities and actually read every document except tabular listings of sports results or public announcements in tabular form. Since the *Frankfurter Rundschau* is also a regional newspaper, articles on nearby cities, towns and villages are frequent. Would one consider the selling of parking meters to another town an event? Or a public invitation to fruit picking or the announcement of a new vocational training as nurse? As the other assessors did not face such a problem, we decided to be rather strict, i.e. an event must be something popular, public and have a certain scope or importance (not only for a single person or a certain group) like concerts, strikes, flooding or sports. In a similar manner, we agreed on a narrower interpretation of the concept of description for GC026 and for GC050 as something unique or characteristic to a city like statistical figures, historical reviews or landmarks. What would be usually considered a description was not often found due to the kind of collection, likewise relevant documents for GC045 *tourism in Northeast Brazil* were also few. While the *SDA* news agency articles will not treat traveling or tourism, such articles may sometimes be found in *Frankfurter Rundschau* or *Spiegel*, but there is no special section on that issue.

Finally, for topic GC027 errors within the documents from the *Frankfurter Rundschau* will have influenced retrieval results: some articles have duplicates (sometimes even up to four versions), different articles thrown together in one document (e.g. one about *Frankfurt* and one about *Wiesbaden*), sentences or passages of articles are missing. Thus a keyword approach may have found many relevant documents, because *Frankfurt* was mentioned somewhere in the document.

5.3 Portuguese Relevance Assessment

Details of Portuguese group’s assessment are as follows: The assessor tried to find the best collection of keywords – based on the detailed information in the narrative and his/her knowledge of the geographical concepts and subjects involved – and queried the DIRECT system. Often there was manual refinement of the query after finding new spellings in previous hits (note that our collections are written in two different varieties of Portuguese). For example, for topic GC050, “cities along the Danube and the Rhine”, the following (final) query was used: *Danúbio Reno Ulm Ingolstadt Regensburg Passau Linz Krems Viena Bratislava Budapeste Vukovar Novi Sad Belgrado Drobeta-Turnu Severin Vidin Ruse Brăila Galați Tulcea Braila Galati Basel Basileia Basileia Estrasburgo Strasbourg Karlsruhe Karlsruhe Mannheim Ludwigshafen Wiesbaden Mainz Koblenz Coblença Bona Bonn Colônia Colônia Cologne*

Düsseldorf Dusseldorf Dusseldórfia Neuss Krefeld Duisburg Duisburgo Arnhem Nederrijn Arnhemia Nijmegen Waal Noviomago Utrecht Kromme Rijn Utreque Rotterdam Roterdão. A similar strategy was used for cities within 100 km from Frankfurt am Main (GC027), where both particular cities were mentioned, as well as words like *cidade* (city), Frankfurt, *distância* (distance), and so on. Obviously, the significant passages for all hits were read, to assess whether the document actually mentioned cities near Frankfurt.

5.4 Spanish Relevance Assessment

For the evaluation of the Spanish documents in the GeoCLEF task, the research group of Language and Information Systems at the University of Alicante, Spain followed the following procedure. The returned documents for each topic from the GeoCLEF collection have been assigned to a member of the group. Each member had to read the question and to identify the relevant keywords. Afterwards, these keywords have been searched in the document together with the geographic names that appeared in the documents. The names were queried in the GeoNames database in order to determine whether this location corresponded to the necessary latitude and magnitude. In addition, the assessors read the title, the narrative and the content of the document, and on the basis of this information decided whether the answer is relevant with the presented topic. The total number of assessors who took part in the evaluation procedure was 36.

5.5 Challenges to Relevance Assessment

One of the major challenges facing the assessors in GeoCLEF 2006 was the substantial increase in the level of geographic knowledge required to assess particular topics. As discussed above, topic GC027 (cities around Frankfurt) was assessed after creating a custom list of cities within 100 km of Frankfurt from the NGA gazetteer. However, for GC050 (cities along the Danube and the Rhine rivers), no equivalent database was created, and the creation of such a database would have posed major challenges. The details of this challenge are described in a separate paper presented at the workshop on Evaluation of Information Access in Tokyo in May 2007 [4].

6 GeoCLEF Performance

6.1 Participants and Experiments

As shown in Table 1, a total of 17 groups from 8 different countries submitted results for one or more of the GeoCLEF tasks - an increase on the 13 participants of last year. A total of 149 experiments were submitted, which is an increase on the 117 experiments of 2005. There is almost no variation in the average number of submitted runs per participant: from 9 runs/participant of 2005 to 8.7 runs/participant of this year.

Table 1. GeoCLEF 2006 participants – new groups are indicated by *

Participant	Institution	Country
alicante	University of Alicante	Spain
berkeley	University of California, Berkeley	United States
daedalus*	Daedalus Consortium	Spain
hagen	University of Hagen	Germany
hildesheim*	University of Hildesheim	Germany
imp-coll*	Imperial College London (imp-coll)*	United Kingdom
jaen*	University of Jaen	Spain
ms-china*	Microsoft China – Web Search and Mining Group	China
nicta	NICTA, University of Melbourne	Australia
rfia-upv	Universidad Politècnica de Valencia	Spain
sanmarcos	California State University, San Marcos	United States
talp	TALP – Universitat Politècnica de Catalunya	Spain
u.buffalo*	SUNY at University of Buffalo	United States
u.groningen*	University of Groningen	The Netherlands
u.twente*	University of Twente	The Netherlands
unsw*	University of New S. Wales	Australia
xldb	Grupo XLDB – Universidade de Lisboa	Portugal

Table 2 reports the number of participants by their country of origin.

Table 2. GeoCLEF 2006 participants by country

Country	# Participants
Australia	2
China	1
Germany	2
Portugal	1
Spain	5
The Netherlands	2
United Kingdom	1
United States	3
TOTAL	17

Table 3 provides a breakdown of the experiments submitted by each participant for each of the offered tasks. With respect to last year there is an increase in the number of runs for the monolingual English task (73 runs in 2006 *wrt* 53 runs of 2005) and a decrease in the monolingual German (16 runs in 2006 *wrt* 25 runs in 2005); on the other hand, there is a decrease for both bilingual English (12 runs in 2006 *wrt* 22 runs in 2005) and bilingual German (11 runs in 2006 *wrt* 17 runs in 2005). Note that the Spanish and the Portuguese collections have been introduced this year.

Table 3. GeoCLEF 2006 experiments by task – new collections are indicated by*

Participant	Monolingual Tasks				Bilingual Tasks				TOTAL
	DE	EN	ES*	PT*	X2DE	X2EN	X2ES*	X2PT*	
alicante		4	3						7
berkeley	2	4	2	4	2		2	2	18
daedalus	5	5	5						15
hagen	5				5				10
hildesheim	4	5			4	5			18
imp-coll		2							2
jaen		5				5			10
ms-china		5							5
nicta		5							5
rfia-upv		4							4
sanmarcos		5	5	4		2	3	2	21
talp		5							5
u.buffalo		4							4
u.groningen		5							5
u.twente		5							5
unsw		5							5
xldb		5		5					10
TOTAL	16	73	15	13	11	12	5	4	149

Four different topic languages were used for GeoCLEF bilingual experiments. As always, the most popular language for queries was English; German and Spanish tied for the second place. Note that Spanish is a new collection added this year. The number of bilingual runs by topic language is shown in Table 4.

Table 4. Bilingual experiments by topic language

Track	Source Language				TOTAL
	DE	EN	ES	PT	
Bilingual X2DE		11			11
Bilingual X2EN	7		5		12
Bilingual X2ES		3		2	5
Bilingual X2PT		2	2		4
TOTAL	7	16	7	2	32

6.2 Monolingual Experiments

Monolingual retrieval was offered for the following target collections: English, German, Portuguese, and Spanish. As can be seen from Table 3, the number of participants and runs for each language was quite similar, with the exception of English, which has the greatest participation. Table 5 shows the top five groups for each target collection, ordered by mean average precision. Note that only the best run is selected

Table 5. Best entries for the monolingual track (title+description topic fields only). Performance difference between the best and the last (up to 5) group is given (in terms of average precision) – new groups indicated by *.

Track		Participant Rank					Diff.
		1 st	2 nd	3 rd	4 th	5 th	
Monolingual English	Part.	xldb	alicante	sanmarcos	unsw*	jaen*	
	Run	XLDBGeoManualEN not pooled	enTD pooled	SMGeoEN4 not pooled	unswTitle-Baseline pooled	sinaiE-nEnExp4 not pooled	
	Avg. Prec.	30.34%	27.23%	26.37%	26.22%	26.11%	16.20%
Monolingual German	Part.	hagen	berkeley	hildesheim*	daedalus*		
	Run	FUHddGY YYTD pooled	BKGeoD1 pooled	HIGeoded-erun4 pooled	GCdeNtLg pooled		
	Avg. Prec.	22.29%	21.51%	15.58%	10.01%		122.68%
Monolingual Portuguese	Part.	xldb	berkeley	sanmarcos			
	Run	XLDBGeoManualPT pooled	BKGeoP3 pooled	SMGeoPT2 pooled			
	Avg. Prec.	30.12%	16.92%	13.44%			124,11%
Monolingual Spanish	Part.	alicante	berkeley	daedalus*	Sanmarcos		
	Run	esTD pooled	BKGeoS1 pooled	GCesNtLg pooled	SMGeoES1 pooled		
	Avg. Prec.	35.08%	31.82%	16.12%	14.71%		138,48%

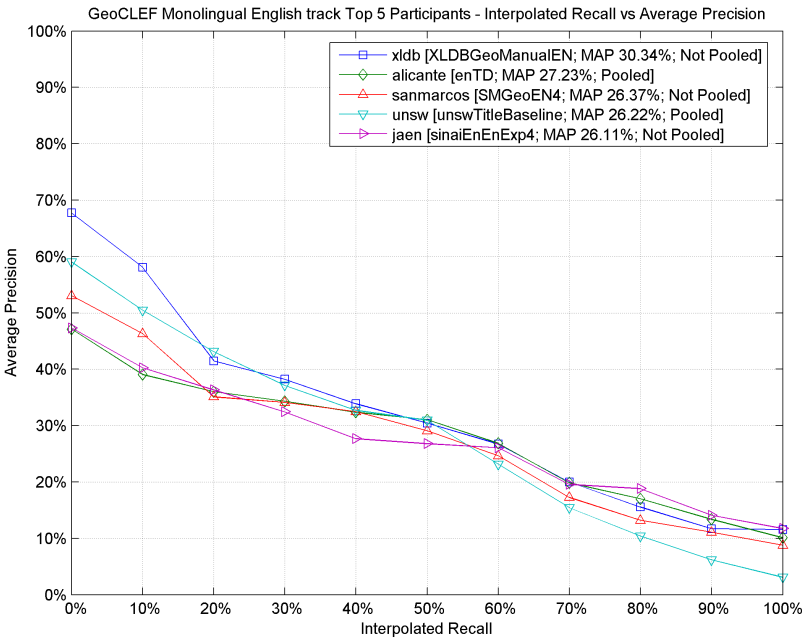


Fig. 2. Monolingual English top participants. Interpolated Recall vs. Average Precision.

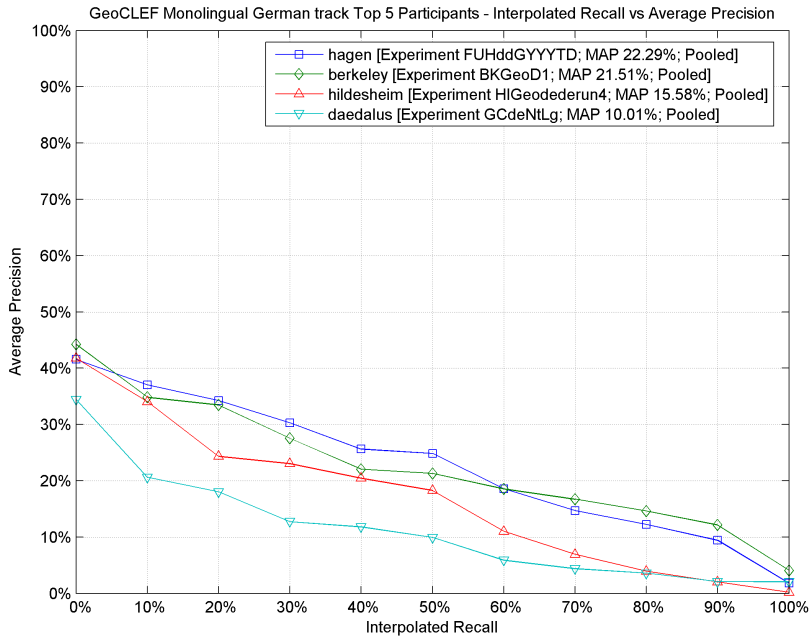


Fig. 3. Monolingual German top participants. Interpolated Recall vs. Average Precision.

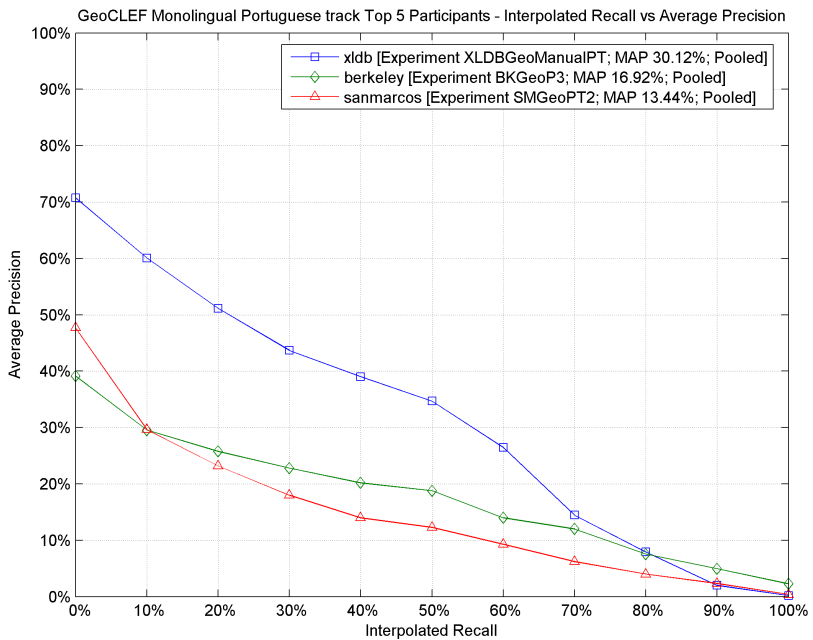


Fig. 4. Monolingual Portuguese top participants. Interpolated Recall vs. Average Precision.

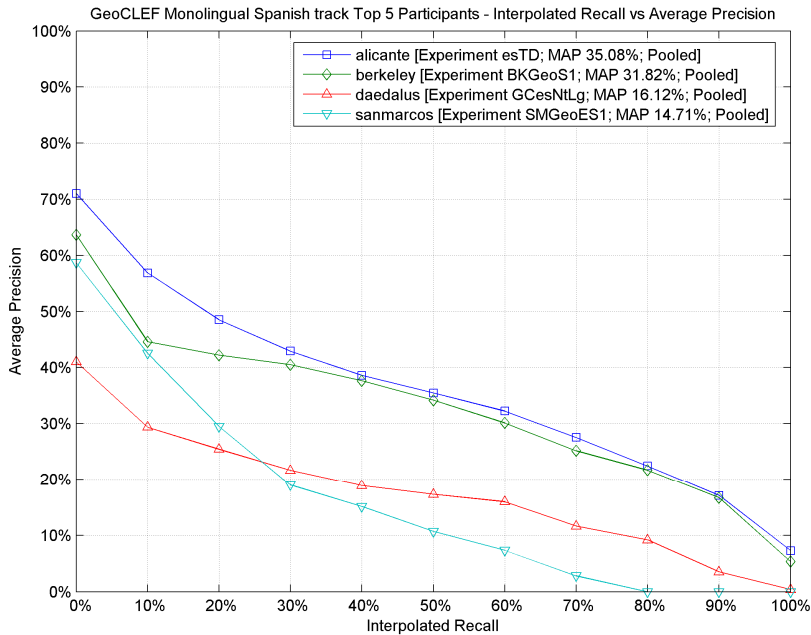


Fig. 5. Monolingual Spanish top participants. Interpolated Recall vs. Average Precision.

for each group, even if the group may have more than one top run. The table reports: the short name of the participating group; the run identifier, specifying whether the run has participated in the pool or not; the mean average precision achieved by the run; and the performance difference between the first and the last participant. Table 5 regards runs using title + description fields only.

Note that the top five participants contain both “newcomer” groups (i.e. groups that had not previously participated in GeoCLEF) and “veteran” groups (i.e. groups that had participated in previous editions of GeoCLEF), with the exception of monolingual Portuguese where only “veteran” groups were subscribed. Both pooled and not pooled runs are in the best entries for each track.

Figures 2 to 5 show the interpolated recall vs. average precision for top participants of the monolingual tasks.

6.3 Bilingual Experiments

The bilingual task was structured in four subtasks ($X \rightarrow$ DE, EN, ES or PT target collection). Table 6 shows the best results for this task with the same logic of Table 5. Note that the top five participants contain both “newcomer” groups and “veteran” groups, with the exception of monolingual Portuguese and Spanish where only “veteran” groups were subscribed.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines:

- $X \rightarrow$ DE: 70% of best monolingual German IR system
- $X \rightarrow$ EN: 74% of best monolingual English IR system

- X → ES: 73% of best monolingual Spanish IR system
- X → PT: 47% of best monolingual Portuguese IR system

Note that the apparently different result for Portuguese may be explained by the fact that the best group in the monolingual experiments did not submit runs for bilingual experiments. If one compares the results per groups, sanmarcos's run of Spanish to Portuguese had even better results than their monolingual Portuguese run, while berkeley's English to Portuguese achieved a similar performance degradation as the one reported for the other bilingual experiments (74%).

Table 6. Best entries for the bilingual task (title+description topic fields only). The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision) – new groups are indicated by *

Track		Participant Rank					
		1 st	2 nd	3 rd	4 th	5 th	Diff.
Bilingual English	Part .	jaen*	sanmarcos	Hildesheim*			
	Run	sinai-ESENEXP 2 pooled	SMGeoE-SEN2 pooled	HIGeodeen-run12 pooled			
	Avg. Prec .	22.56%	22.46%	16.03%			40.74%
Bilingual German	Part .	berkeley	hagen	Hildesheim*			
	Run	BKGeoED 1 pooled	FU-HedGY-YYTD pooled	HI-Geoenderun21 pooled			
	Avg. Prec .	15.61%	12.80%	11.86%			31.62%
Bilingual Portuguese	Part .	sanmarcos	berkeley				
	Run	SMGeoES PT2 pooled	BKGeoEP 1 pooled				
	Avg. Prec .	14.16%	12.60%				12,38%
Bilingual Spanish	Part .	berkeley	sanmarcos				
	Run	BKGeoES 1 pooled	SMGeoE NES1 pooled				
	Avg. Prec .	25.71%	12.82%				100.55%

Figure 6 to 9 show the interpolated recall vs. average precision graph for the top participants of the different bilingual tasks.

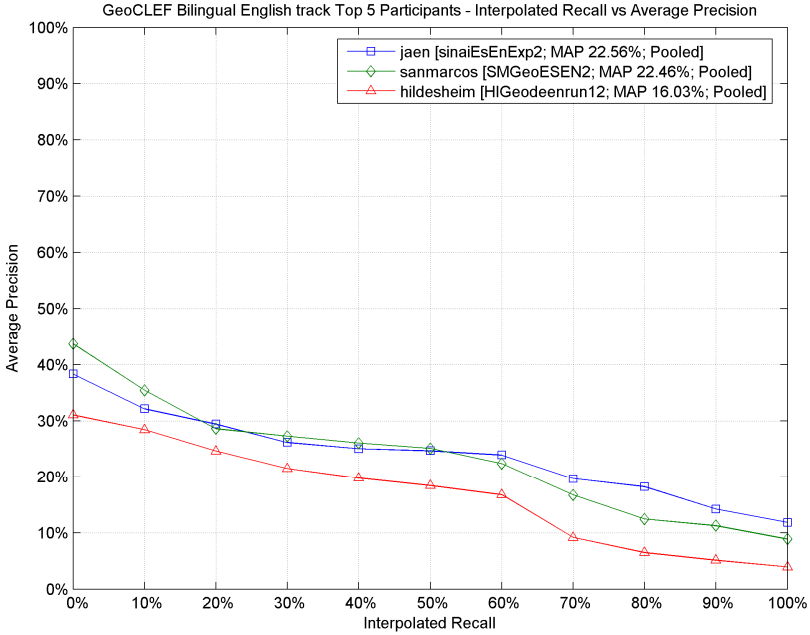


Fig. 6. Bilingual English top participants. Interpolated Recall vs Average Precision.

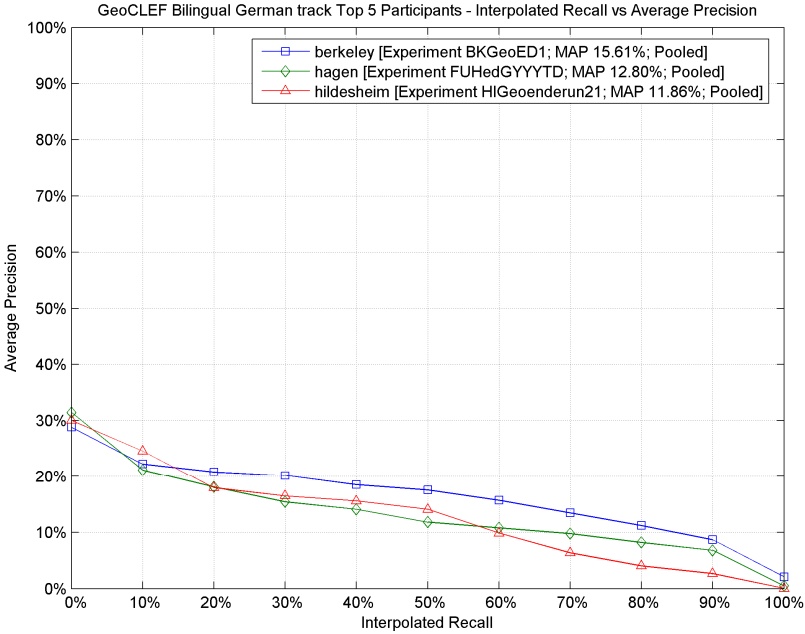


Fig. 7. Bilingual German top participants. Interpolated Recall vs Average Precision.

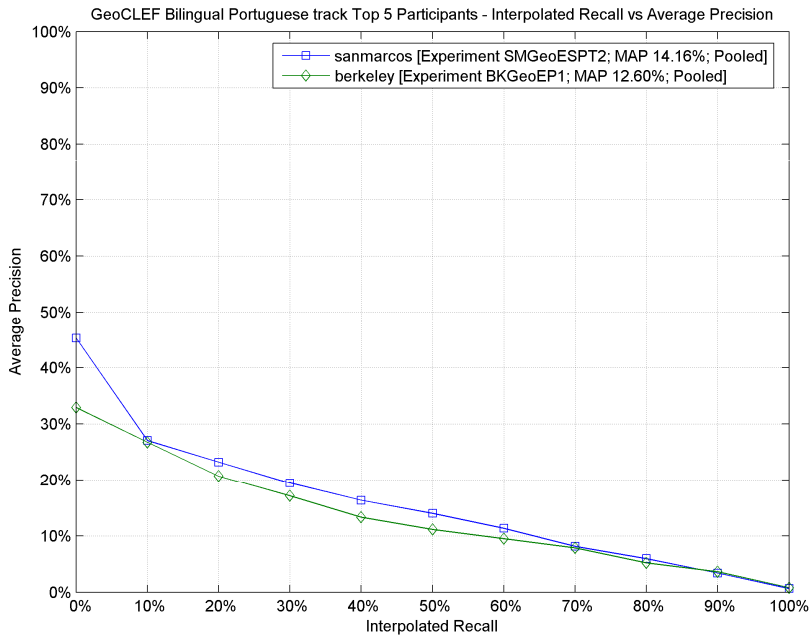


Fig. 8. Bilingual Portuguese top participants. Interpolated Recall vs Average Precision.

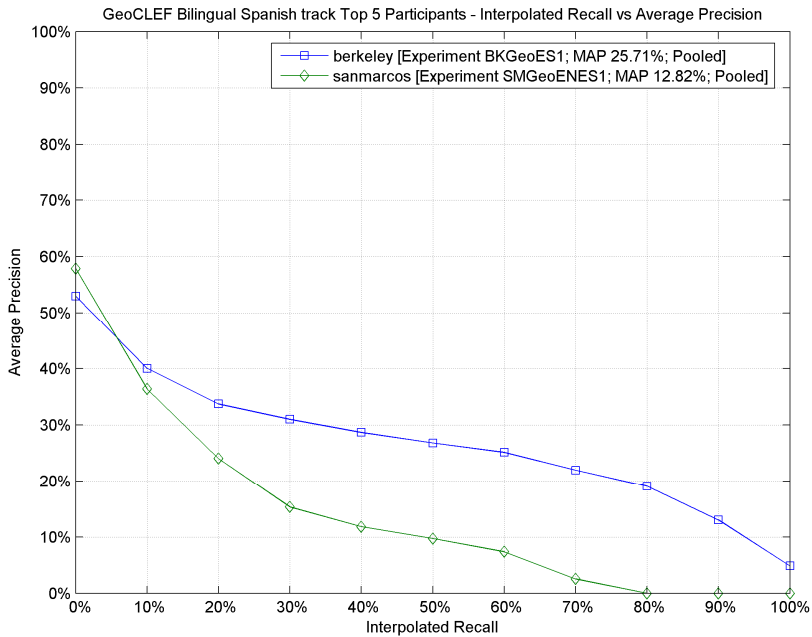


Fig. 9. Bilingual Spanish top participants. Interpolated Recall vs Average Precision.

6.4 Statistical Testing

We used the MATLAB Statistics Toolbox, which provides the necessary functionality plus some additional functions and utilities. We use the *ANalysis Of VAriance* (ANOVA) test. ANOVA makes some assumptions concerning the data be checked. Hull [5] provides details of these; in particular, the scores in question should be approximately normally distributed and their variance has to be approximately the same for all runs. Two tests for goodness of fit to a normal distribution were chosen using the MATLAB statistical toolbox: the Lilliefors test [1] and the Jarque-Bera test [6]. In the case of the GeoCLEF tasks under analysis, both tests indicate that the assumption of normality is violated for most of the data samples (in this case the runs for each participant).

In such cases, a transformation of data should be performed. The transformation for measures that range from 0 to 1 is the arcsin-root transformation:

$$\arcsin(\sqrt{x})$$

which Tague-Sutcliffe [11] recommends for use with precision/recall measures.

Table 7. Lilliefors test for each track with (LL) and without Tague-Sutcliffe arcsin transformation (LL & TS). Jarque-Bera test for each track with (JB) and without Tague-Sutcliffe arcsin transformation (JB & TS).

Track	LL	LL & TS	JB	JB & TS
Monolingual English	2	42	32	54
Monolingual German	0	3	3	11
Monolingual Portuguese	4	5	3	13
Monolingual Spanish	2	12	4	11
Bilingual English	0	4	2	6
Bilingual German	0	1	0	4
Bilingual Portuguese	0	3	0	4
Bilingual Spanish	0	2	2	5

Table 7 shows the results of the Lilliefors test before and after applying the Tague-Sutcliffe transformation. After the transformation the analysis of the normality of samples distribution improves significantly, with the exception of the bilingual Bulgarian. Each entry shows the number of experiments whose performance distribution can be considered drawn from a Gaussian distribution, with respect to the total number of experiment of the track. The value of alpha for this test was set to 5%. The same table shows also the same analysis with respect to the Jarque-Bera test. The value of alpha for this test was set to 5%. The difficulty to transform the data into normally distributed samples derives from the original distribution of run performances which tend towards zero within the interval [0,1].

The following tables, from Table 8 to Table 13, summarize the results of this test. All experiments, regardless the topic language or topic fields, are included. Results are therefore only valid for comparison of individual pairs of runs, and not in terms of

absolute performance. Each table shows the overall results where all the runs that are included in the same group do not have a significantly different performance. All runs scoring below a certain group performs significantly worse than at least the top entry of the group. Likewise all the runs scoring above a certain group perform significantly better than at least the bottom entry in that group. Each table contains also a graph which shows participants' runs (y axis) and performance obtained (x axis). The circle

Table 8. Monolingual English: experiment groups according to the Tukey T Test

Run ID	Groups									
XLDBGeoManualEN	X									
sinaiEnEnExp1	X									
enTDN	X	X								
SMGeoEN3	X	X								
unswNarrBaseline	X	X								
BKGeoE4	X	X								
BKGeoE3	X	X	X							
SMGeoEN1	X	X	X							
SMGeoEN4	X	X	X							
unswTitleBaseline	X	X	X							
sinaiEnEnExp4	X	X	X	X						
enTD	X	X	X	X						
rfialUPV02	X	X	X	X						
BKGeoE2	X	X	X	X	X					
rfialUPV04	X	X	X	X	X					
sinaiEnEnExp2	X	X	X	X	X					
BKGeoE1	X	X	X	X	X					
MuTdnTxt	X	X	X	X	X	X				
sinaiEnEnExp5	X	X	X	X	X	X				
rfialUPV01	X	X	X	X	X	X				
UBManual2	X	X	X	X	X	X				
SMGeoEN5	X	X	X	X	X	X				
MuTdnTxt	X	X	X	X	X	X				
UBGTDrf1	X	X	X	X	X	X				
UBGTDrf2	X	X	X	X	X	X				
msramanual	X	X	X	X	X	X				
MuTdnManOexpGeo	X	X	X	X	X	X				
sinaiEnEnExp3	X	X	X	X	X	X				
UBGManual1	X	X	X	X	X	X				
MuTdnRedn	X	X	X	X	X	X				
rfialUPV03	X	X	X	X	X	X				
unswTitleF46	X	X	X	X	X	X				
UAUJAUPVenenExp1	X	X	X	X	X	X				
XLDBGeoENAut05	X	X	X	X	X	X				
MuTdnOexpPrb	X	X	X	X	X	X				
CLCGGeoEE11	X	X	X	X	X	X	X			
CLCGGeoEE2	X	X	X	X	X	X	X			
XLDBGeoENAut03_2	X	X	X	X	X	X	X			
msrawhitelist	X	X	X	X	X	X	X			
ICgeoMLtdn	X	X	X	X	X	X	X			
msralocal	X	X	X	X	X	X	X			
msraltxt	X	X	X	X	X	X	X			
CLCGGeoEE1	X	X	X	X	X	X	X			
utGeoTIBm	X	X	X	X	X	X	X			
utGeoTdnIBm	X	X	X	X	X	X	X			
XLDBGeoENAut03	X	X	X	X	X	X	X			
CLCGGeoEE5	X	X	X	X	X	X	X			
utGeoTIB	X	X	X	X	X	X	X			
CLCGGeoEE10	X	X	X	X	X	X	X			
XLDBGeoENAut02	X	X	X	X	X	X	X			
HIGeoenenrun3	X	X	X	X	X	X	X			
ICgeoMLtd	X	X	X	X	X	X	X			
HIGeoenenrun1n	X	X	X	X	X	X	X			
msrlexpansion	X	X	X	X	X	X	X			
HIGeoenenrun1	X	X	X	X	X	X	X			
GCenATLg	X	X	X	X	X	X	X			
TALPGeoIRTD1	X	X	X	X	X	X	X			
TALPGeoIRTD1	X	X	X	X	X	X	X			
GCenAA	X	X	X	X	X	X	X			
HIGeoenenrun2n	X	X	X	X	X	X	X			
utGeoTdnIB	X	X	X	X	X	X	X			
enTDNGeoNames	X	X	X	X	X	X	X			
GCenNLg	X	X	X	X	X	X	X			
TALPGeoIRTD3	X	X	X	X	X	X	X			
HIGeoenenrun2	X	X	X	X	X	X	X			
GCenAO	X	X	X	X	X	X	X			
utGeoTdnIB	X	X	X	X	X	X	X			
GCenNA	X	X	X	X	X	X	X			
TALPGeoIRTD2	X	X	X	X	X	X	X			
unswNarrF41	X	X	X	X	X	X	X			
TALPGeoIRTD2	X	X	X	X	X	X	X			
unswNarrMap	X	X	X	X	X	X	X			

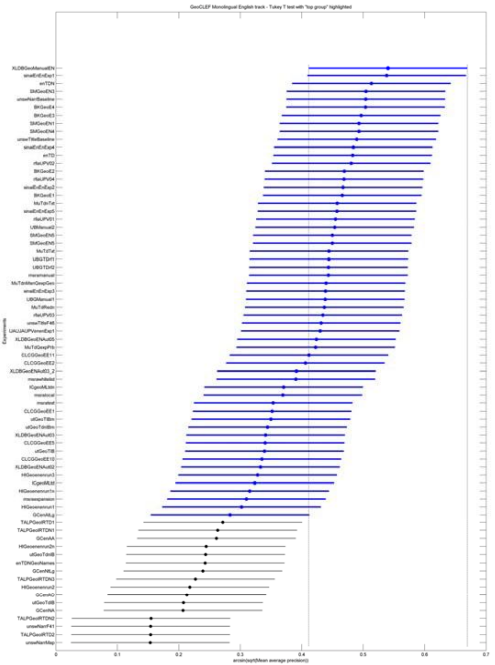


Table 9. Monolingual German: experiment groups according to the Tukey T Test

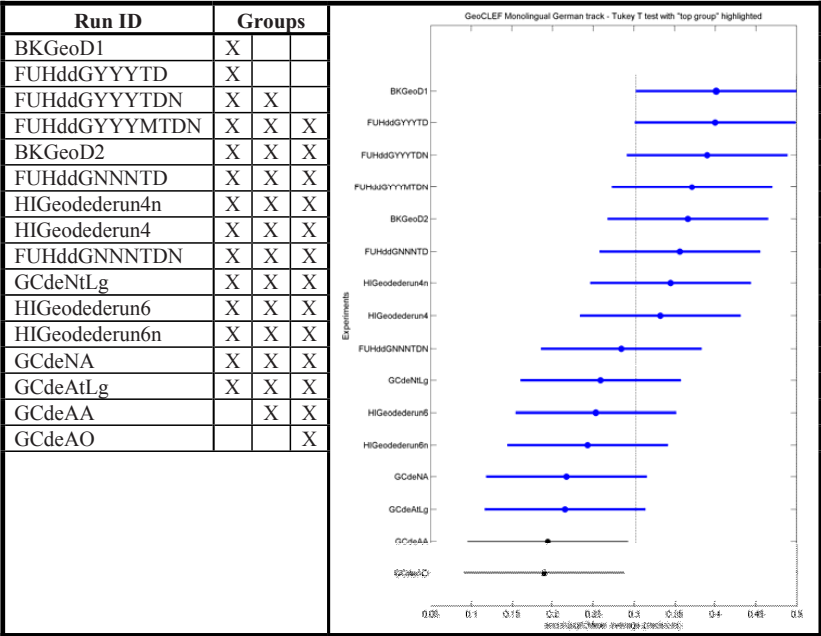


Table 10. Monolingual Portuguese: experiment groups according to the Tukey T Test

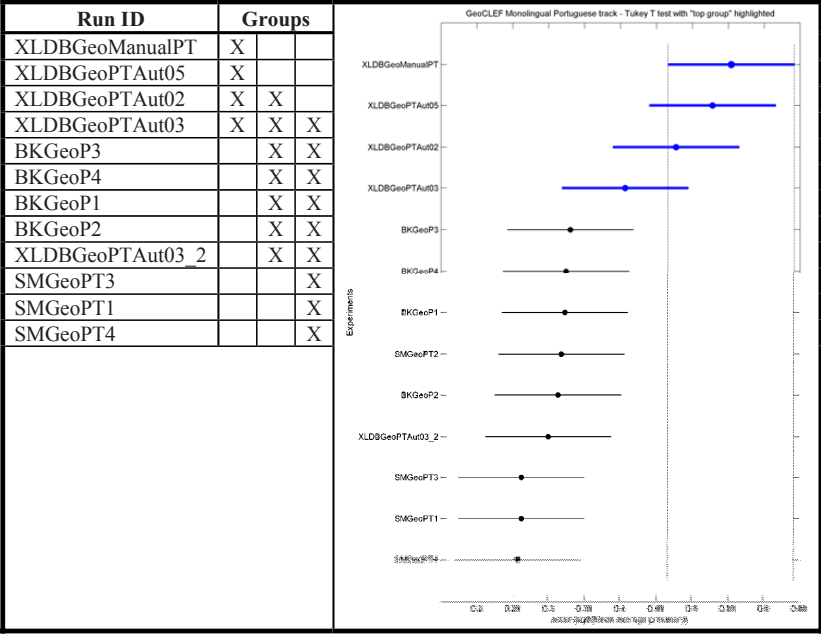


Table 11. Monolingual Spanish: experiment groups according to the Tukey T Test

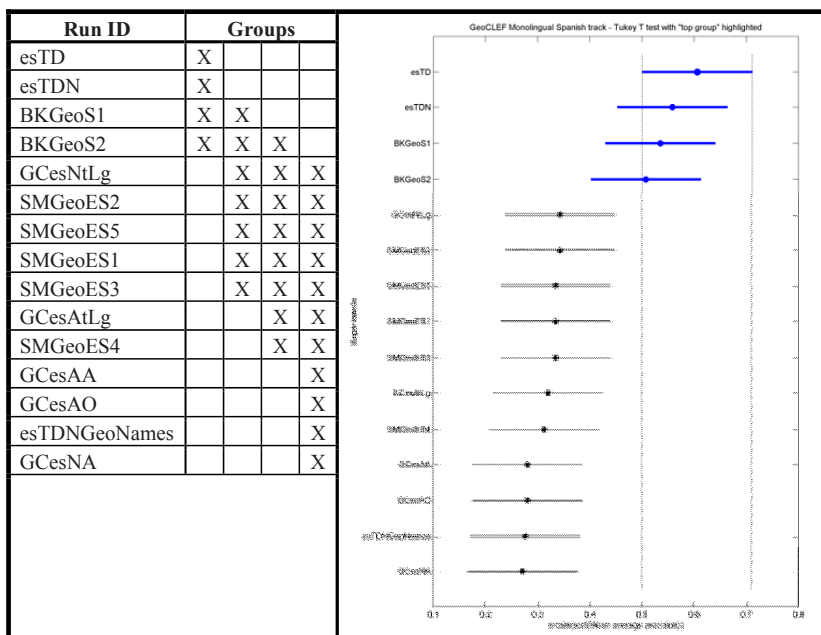


Table 12. Bilingual English: experiment groups according to the Tukey T Test

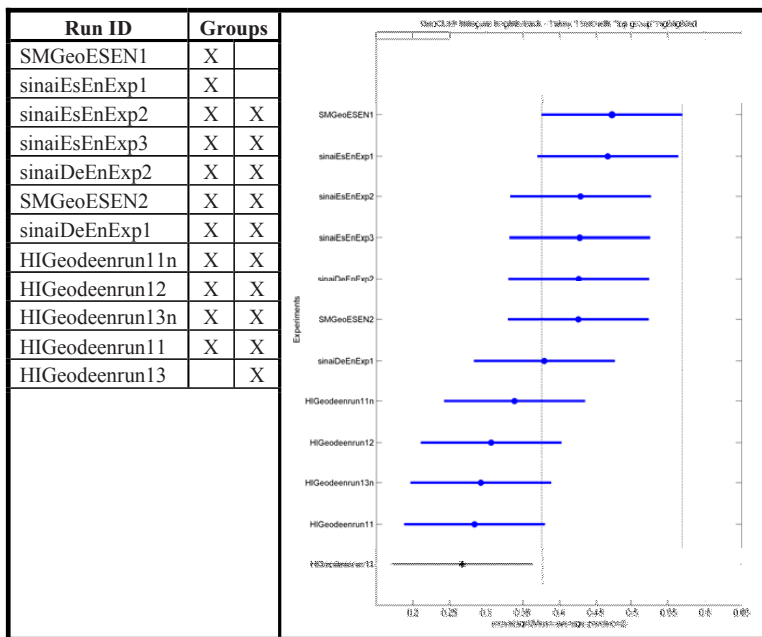
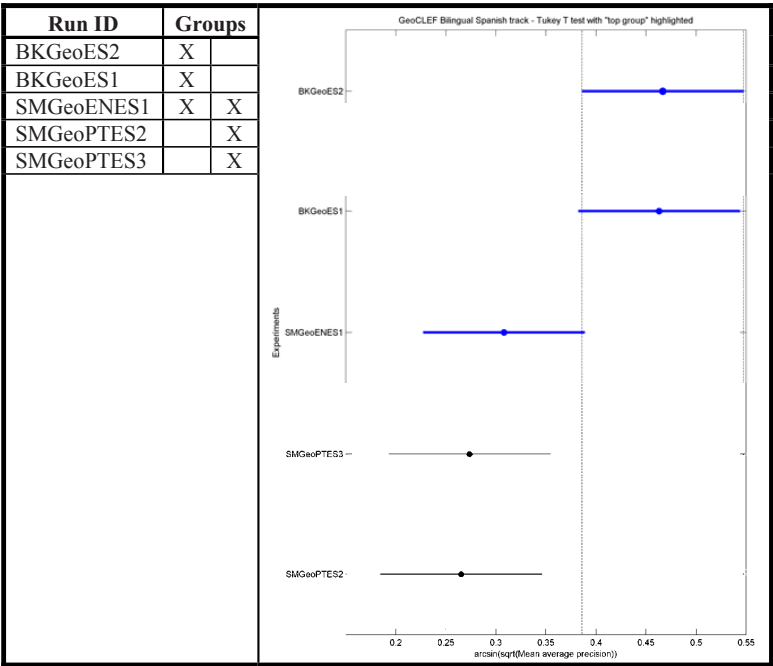


Table 13. Bilingual Spanish: experiment groups according to the Tukey T Test



indicates the average performance while the segment shows the interval in which the difference in performance is not statistically significant; for each graph the best group is highlighted.

Note that there are no tables for Bilingual German and Bilingual Portuguese since, according to the Tukey T, all the experiments of these tasks belong to the same group.

6.5 Limitations to the Non-english Experiments, Particularly Portuguese

One must be cautious about drawing conclusions about system performance where only a limited number of groups submitted experimental runs. For the non-English collections this is especially true for GeoCLEF 2006. Six groups submitted German language runs, four groups submitted Spanish runs, and three groups submitted Portuguese runs. The Portuguese results are dominated by the XLDB group (the only participating group with native Portuguese language speakers). XLDB’s overall results are nearly twice that of Berkeley’s and more than twice that of San Marcos. XLDB contributed uniquely nearly 40% of the relevant Portuguese documents found, most of them through their manual run. Thus the Portuguese results of the other groups would have been higher without the XLDB participation.

7 Conclusions and Future Work

GeoCLEF 2006 increased the geographic challenge contained within the topics over 2005. Several topics could not be fully exploited without external gazetteer resources. Such resources were utilized by both the participants and the organizing groups doing relevance assessment. The test collection developed for GeoCLEF is the first GIR test collection available to the GIR research community. GIR is receiving increased notice both through the GeoCLEF effort as well as due to the GIR workshops held annually since 2004 in conjunction with SIGIR or CIKM.

At the GIR06 workshop held in August 2006, in conjunction with SIGIR 2006 in Seattle, 14 groups participated and 16 full papers were presented, as well as a keynote address by John Frank of MetaCarta, and a summary of GeoCLEF 2005 presented by Ray Larson. Six of the groups also were participants in GeoCLEF 2005 or 2006. Six of the full papers presented at the GIR workshop used GeoCLEF collections. Most of these papers used the 2005 collection and queries, although one group used queries from this year's collection with their own take on relevance judgments. Of particular interest to the organizers of GeoCLEF are the 8 groups working in the area of GIR who are not yet participants in GeoCLEF. All attendees at the GIR06 workshop were invited to participate in GeoCLEF for 2007.

Acknowledgments

The English assessment by the GeoCLEF organizers was volunteer labor – none of us has funding for GeoCLEF work. Assessment was performed by Hans Barnum, Nils Bailey, Fredric Gey, Ray Larson, and Mark Sanderson. German assessment was done by Claudia Bachmann, Kerstin Bischoff, Thomas Mandl, Jens Plattfaut, Inga Rill and Christa Womser-Hacker of University of Hildesheim. Portuguese assessment was done by Paulo Rocha, Luís Costa, Luís Cabral, Susana Inácio, Ana Sofia Pinto, António Silva and Rui Vilela, all of Linguatca, and Spanish by Andrés Montoyo Guijarro, Oscar Fernandez, Zornitsa Kozareva, Antonio Toral of University of Alicante. The Linguatca work on Portuguese was supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI. The GeoCLEF website is maintained by University of Sheffield, England, United Kingdom. The future direction and scope of GeoCLEF will be heavily influenced by funding and the amount of volunteer effort available.

References

- [1] Conover, W.J.: Practical Nonparametric Statistics. John Wiley and Sons, New York, USA (1971)
- [2] Eilders, C.: The role of news factors in media use. Technical report, Wissenschaftszentrum Berlin für Sozialforschung gGmbH (WZB). Forschungsschwerpunkt Sozialer Wandel, Institutionen und Vermittlungsprozesse des Wissenschaftszentrums Berlin für Sozialforschung. Berlin (1996)

- [3] Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giamppiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [4] Gey, F., et al.: Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF. In: Proceedings of the 1st International Workshop on Evaluation of Information Access, Tokyo Japan, May 15, 2007, pp. 74–77 (2007)
- [5] Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Korfhage, R., Rasmussen, E., Willett, P. (eds.) Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1993), pp. 329–338. ACM Press, New York (1993)
- [6] Judge, G.G., Hill, R.C., Griffiths, W.E., Lutkepohl, H., Lee, T.C.: Introduction to the Theory and Practice of Econometrics, 2nd edn. John Wiley and Sons, New York, USA (1988)
- [7] Kluck, M., Womser-Hacker, C.: Inside the evaluation process of the cross-language evaluation forum (CLEF): Issues of multilingual topic creation and multilingual relevance assessment. In: Proceedings of the third International Conference on Language Resources and Evaluation, LREC 2002, Las Palmas, Spain, pp. 573–576 (2002)
- [8] Santos, D., Bick, E.: Providing internet access to portuguese corpora: the ac/dc project. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhauer, G. (eds.) Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, Athens, 31 May–2 June 2000, pp. 205–210 (2000)
- [9] Santos, D., Cardoso, N.: Portuguese at CLEF 2005: Reflections and challenges. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giamppiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
- [10] Santos, D., Chaves, M.: The place of place in geographical information retrieval. In: Jones, C., Purves, R. (eds.) Workshop on Geographic Information Retrieval (GIR06), SIGIR06, Seattle, August 10, 2006, pp. 5–8 (2006)
- [11] Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: Sparck, K., Willett, P. (eds.) Readings in Information Retrieval, pp. 205–216. Morgan Kaufmann Publishers, San Francisco, California (1997)

MIRACLE's Ad-Hoc and Geographical IR Approaches for CLEF 2006

José M. Goñi-Menoyo¹, José C. González-Cristóbal², Sara Lana-Serrano¹,
and Ángel Martínez-González²

¹ Universidad Politécnica de Madrid

² DAEDALUS – Data, Decisions, and Language S.A.
josemiguel.goni@upm.es, jgonzalez@dit.upm.es,
slana@diatel.upm.es, amartinez@daedalus.es

Abstract. This paper presents the 2006 Miracle team's approaches to the Ad-Hoc and Geographical Information Retrieval tasks. A first set of runs was obtained using a set of basic components. Then, by putting together special combinations of these runs, an extended set was obtained. With respect to previous campaigns some improvements have been introduced in our system: an entity recognition prototype is integrated in our tokenization scheme, and the performance of our indexing and retrieval engine has been improved. For GeoCLEF, we tested retrieving using geo-entity and textual references separately, and then combining them with different approaches.

1 Introduction

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a company founded in 1998 as a spin-off of two of these groups. DAEDALUS is a leading company in linguistic technologies in Spain and is the coordinator of the MIRACLE team. This is our fourth participation in CLEF since 2003. As well as bilingual, monolingual and robust multilingual tasks, the team has participated in the ImageCLEF, Q&A, and GeoCLEF tracks.

For this campaign, runs were submitted for the following languages and tracks:

- Ad-hoc monolingual: Bulgarian, French, Hungarian, and Portuguese.
- Ad-hoc bilingual: English to Bulgarian, French, Hungarian, and Portuguese; Spanish to French and Portuguese; and French to Portuguese.
- Ad-hoc robust monolingual: German, English, Spanish, French, Italian, and Dutch.
- Ad-hoc robust bilingual: English to German, Italian to Spanish, and French to Dutch.
- Ad-hoc robust multilingual: English to robust monolingual languages.
- Geo monolingual: English, German, and Spanish.

2 MIRACLE Toolbox

All document collections and topic files are processed before feeding the indexing and retrieval engine. This processing is carried out using different combinations of

elementary components. The details can be consulted in the papers from 2006 workshop [2], [3].

These components include (i) extraction, which incorporates a special process to filter out some sentences from the topic narrative that matches a number of recurrent and misleading patterns; (ii) tokenization, which extracts basic text components, such as single words, years, or entities; (iii) entity detection, integrated in the tokenization module, and having a central role in IR processes: for now, it detects previously collected entities and integrates them into a special resource; (iv) filtering, which eliminates stopwords and words without semantic content in the CLEF context; (v) transformation, which normalizes case and diacritics; (vi) stemming [4], [6]; (vii) indexing, using our own trie-based [1] tool; and (viii) retrieval engine, which implements the well-known Robertson's Okapi [5] BM-25 formula for probabilistic retrieval model, without relevance feedback.

After retrieval, a number of other special combination processes were used to define additional experiments. The results from several basic experiments are combined using two different strategies: average and asymmetric WDX combination (see [2]). The underlying hypothesis for these combinations is that highly scored documents in several experiments are more likely to be relevant than other documents that have good scores in some experiments but bad ones in others.

We used the traditional approach to multilingual information retrieval that translates topic queries into the target language of the document collections. The probabilistic BM25 approach used for monolingual retrieval gives relevance measures that depends heavily on parameters that are too dependent on the monolingual collection, so it is not very good for this type of multilingual merging, since relevance measures are not comparable between collections. In spite of this, we carried out merging experiments using the relevance figures obtained from each monolingual retrieval process, considering three cases (see [2]): (i) using original relevance measures for each document as obtained from the monolingual retrieval process; (ii) normalizing relevance measures with respect to the maximum relevance measure obtained for each topic query i (*standard normalization*); and (iii) normalizing relevance measures with respect to the maximum and minimum relevance measure obtained for each topic query i (*alternate normalization*). In the three cases, documents with higher resulting relevance are selected from all monolingual results lists. Round-Robin merging for results of each monolingual collection has not been used.

In addition to all this, we tried a different approach to multilingual merging: Considering that the more relevant documents for each of the topics are usually the first ones in the results list, we select a variable number of documents, proportional to the average relevance number of the first N documents from each monolingual results file. Thus, if we need 1,000 documents for a given topic query, we get more documents from languages where the average relevance of the first N relevant documents is greater. We implemented this process in two ways: The appropriate number of documents to be aggregated is computed from both non normalized and normalized runs (using the two normalization formulae).

For Geographical IR, a Gazetteer that drives a named Geo-entity identifier was built as well as a tagger [3]. The gazetteer is a list of geographical resources,

compiled from two existing gazetteers, GNIS and NGA, which required the development of a geographical ontology.

3 Ad-Hoc Experiments

Both in the monolingual and the bilingual cases, the results obtained for “related” languages, such as French and Portuguese, are better than those obtained for Bulgarian and Hungarian. In the bilingual case, French experiments have best average precision.

In all cases, the best results are obtained from the experiments that take the topic narrative into account. Unfortunately, the official published reports only consider experiments that use topic title and description exclusively.

In the robust monolingual case, results for Spanish are much better than those obtained for the other the languages. In all cases the use of baseline runs has obtained better results than the use of combined ones. Curiously, the Dutch target language runs have better results than those in other languages. Note that in all cases, the experiments taking the topic narrative into account show the best results, as happened in the non-robust case.

4 Geographical Retrieval Experiments

The main objective of us in this campaign was to test the effects of geographical IR in documents containing geographical tags. We designed experiments to try to isolate geographical retrieval from textual retrieval. We have replaced all geo-entity textual references with associated tags in each topic, and then we searched all documents for these tags. This is done sequentially by combining a Geo-query Identifier process, a Spatial Relation identifier, and an Expander. These results are combined with those obtained using the usual ad-hoc text retrieval process. For combinations, several techniques were used: union (OR), intersection (AND), difference (AND NOT), and external join (LEFT JOIN). These techniques re-rank the output results by computing new relevance measure values for each document.

5 Conclusions

We still need to work harder to improve a number of aspects of our processing scheme, entity recognition and normalization being the most important ones. It is clear that the quality of the tokenization step is of paramount importance for precise document processing. A high-quality entity recognition (proper nouns or acronyms for people, companies, countries, locations, and so on) could improve the precision and recall figures of the overall retrieval, as well as a correct recognition and normalization of dates, times, numbers, etc. Although we have introduced some improvements to our processing scheme, a good multilingual entity recognition and normalization tool is still missing both for Ad-hoc and Geo IR.

We are also improving the architecture of our indexing and retrieval *trie*-based engine in order to get an even better performance in the indexing and retrieval phases,

tuning some data structures and algorithms. We are now implementing pseudo-relevance feedback and document filtering modules.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, through the RIMMEL project (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01; and by Madrid's R+D Regional Plan, through the MAVIR project (Enhancing the Access and the Visibility of Networked Multilingual Information for Madrid Community), S-0505/TIC/000267.

Special mention of our colleagues from the MIRACLE team should be made (in alphabetical order): Ana María García-Serrano, Ana González-Ledesma, José M^a Guirao-Miras, José Luis Martínez-Fernández, Paloma Martínez-Fernández, Antonio Moreno-Sandoval and César de Pablo-Sánchez.

References

1. Aoe, J.-I., Morimoto, K., Sato, T.: An Efficient Implementation of Trie Structures. *Software Practice and Experience* 22(9), 695–721 (1992)
2. Goñi-Menoyo, J.M., González, J.C., Villena-Román, J.: Report of MIRACLE Team for the Ad-Hoc Track in CLEF. In: *Working Notes for the CLEF 2006 Workshop*. Alicante, Spain (2006) <http://www.clef-camp.aign.org/>
3. Lana-Serrano, S., Goñi-Menoyo, J.M., González, J.C.: Report of MIRACLE Team for Geographical IR in CLEF, *Working Notes for the CLEF 2006 Workshop*. Alicante, Spain, (2006), <http://www.clef-campaign.org/>
4. Porter, M.: Snowball stemmers and resources page. [Visited 30/09/2006] <http://www.snowball.tartarus.org>
5. Robertson, S.E., et al.: Okapi at TREC-3. In: Harman, D.K. (ed.) *Overview of the Third Text REtrieval Conference (TREC-3)*, Gaithersburg, MD: NIST (1995)
6. University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers..). [Visited 30/09/2006] On line <http://www.unine.ch/info/clef>

GIR Experimentation

Andogah Geoffrey

Computational Linguistics Group,
Centre for Language and Cognition Groningen (CLCG),
University of Groningen, Groningen, The Netherlands
`g.andogah@rug.nl`, `annageof@yahoo.com`

Abstract. The Geographic Information Retrieval (GIR) community has generally accepted the thesis that both thematic and geographic aspects of documents are useful for GIR. This paper describes an experiment exploring this thesis by separately indexing and searching geographical relevant-terms (place names, geo-spatial relations, geographic concepts and geographic adjectives). Two indexes were created – one for extracted geographic relevant-terms (footprint document) and one for reference document collections. Footprint and reference document scores are combined by a linear interpolation to obtain an overall score for document relevance ranking. Experimentation with geographic query expansion provided no significant improvement though it stabilized the query result.

1 Introduction

Geographic Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Geographic information can be found in many documents, and therefore, geographic references may be important for Information Retrieval (IR) (e.g. to retrieve, re-rank and visualize search results based on a spatial dimension). Additionally, many documents contain geographic references expressed in multiple languages which may or may not be the same as the query language [1]. The GIR community has accepted the thesis that both thematic (non-geographic aspect) and geographic aspects of documents are important for successful implementation of GIR system. To address this thesis we derive for each reference document in the collection a corresponding footprint document containing place names (e.g. Uganda), geo-spatial relations (e.g. west of), geographic concepts (e.g. country) and geographic adjectives (e.g. Ugandan). The footprint document and reference document were separately indexed and searched. Footprint and reference document scores are combined by a linear interpolation to obtain an overall score for document relevance ranking. Experimentation with geographic query expansion provided no significant improvement though it stabilizes the query result.

2 Approach

We are participating in GeoCLEF evaluation track at CLEF 2006 [2] for the first time. The main motivation for our participation is to experiment with both non-geographic and geographic aspects of a document for GIR as a baseline for further investigation. In this section we describe our approach and resources used. Our approach borrows techniques from ([3], [4], [6], [7]) with a few improvements such as the creation of an index of the footprint document in addition to reference document index, thereby combining query results of the two index searches using linear interpolation.

2.1 Resources

Geographic Knowledge Base. We used the World Gazetteer¹, GONet Names Server² (GNS), Wikipedia³ and WordNet⁴ as the bases for our Geographic Knowledge Base (GKB) for several reasons: free availability, multilingual, coverage of most popular and major places, etc.

Geographic Tagger. Alias-I LingPipe⁵ was used to detect named entities (location, person and organisation), geographic concepts (continent, region, country, city, town, village, etc.), spatial relations (near, in, south of, north west, etc.) and locative adjectives (e.g. Ugandan).

Lucene Search Engine. Apache Lucene⁶ is a high-performance, full-featured text search engine library written entirely in Java. Lucene's default similarity measure is derived from the vector space model (VSM). Lucene was used to index and search both indexes.

2.2 Document Indexing

Document Pre-processing. Documents were pre-processed using Alias-I LingPipe to detect place names (e.g. Kampala), geographic concepts (e.g. city), spatial relations (e.g. west of) and adjectives referring to things or people or language connected to a place (e.g. Ugandan). Instead of performing explicit reference resolution, we argue that all geo-relevant terms detected in a document will some how relate or point to a specific geographic scope or geographic concept in the document's discourse.

Footprint Document Indexing. The footprint documents derived from reference documents contain geo-relevant terms such as place names, geographic

¹ <http://www.world-gazetteer.com>

² <http://earthinfo.nga.mil/gns/html>

³ <http://www.wikipedia.org>

⁴ <http://wordnet.princeton.edu>

⁵ <http://alias-i.com/lingpipe>

⁶ <http://jakarta.apache.org/lucene>

concepts, spatial relations, locative adjectives plus their respective term frequency. The derived footprint documents are indexed using Lucene (see Table 1 for details). Geo-relevant term frequency is factored into the index by adding the same geo-relevant term to the index the number of times it occurs in the document.

Table 1. Footprint document index structure

Field	Lucene Type	Description
docid	Field.Keyword	Document unique identification
place-name	Field.Keyword	Place name e.g. Okollo
concept	Field.Keyword	Geographic concept e.g. county
relation	Field.Keyword	Geographic relation e.g. northwest
adjective	Field.Keyword	Geographic adjective e.g. Ugandan

Reference Document Indexing. The reference document collection provided for experimentation was indexed using Lucene. Document HEADLINE and TEXT contents were combined to create document content for indexing (see Table 2 for details).

Table 2. Reference document index structure

Field	Lucene Type	Description
docid	Field.Keyword	Document unique identification
content	Field.Unstored	Combination of HEADLINE and TEXT tag content

2.3 Linear Interpolation Similarity Score Formula

Lucene Similarity Score Formula. As mentioned in Sect. 2.1, our GIR is based on the Apache Lucene IR library. The Lucene similarity score formula combines several factors to determine the document score for a query [8]:

$$LuceneSim(q, d) = \sum_{t \in q} tf(t \text{ in } d) \cdot idf(t) \cdot boost(t.field \text{ in } d) \cdot lengthNorm(t.field \text{ in } d) \tag{1}$$

where, $tf(t \text{ in } d)$ is the term frequency factor for term t in document d , $idf(t)$ is the inverse document frequency of term t , $boost(t.field \text{ in } d)$ is the field boost set during indexing and $lengthNorm(t.field \text{ in } d)$ is the normalization value of a field given the number of terms in the field.

Interpolated Similarity Score Formula. The interpolated similarity score formula is directly derived from the Lucene similarity score.

$$\begin{aligned} \text{InterpolatedSim}(q, d) &= \lambda_T \text{LuceneSim}(q_T, d_R) + \lambda_G \text{LuceneSim}(q_G, d_F) \\ \lambda_T + \lambda_G &= 1 \end{aligned} \quad (3)$$

where d_R is the reference document, d_F the footprint document, q_T the thematic (non-geographic plus geographic terms) query, q_G the geographic (geo-relevant terms) query, λ_T the thematic interpolation factor and λ_G the geographic interpolation factor. For this work, $\text{LuceneSim}(q_T, d_R)$ and $\text{LuceneSim}(q_G, d_F)$ compute the *Thematic Similarity Score* and *Geographic Similarity Score* respectively.

2.4 Query Formulation for Official GeoCLEF Runs

This section describes the University of Groningen official runs for GeoCLEF 2006. In particular we describe which topic components are used for query formulation and which indexes were used with linear combination for retrieval ranking.

CLCGGeoEE1 and CLCGGeoEE2. These are the GeoCLEF 2006 [2] mandatory runs 1 and 2 queries formulated by topic TITLE-DESC (TD) and TITLE-DESC-NARR (TDN) contents respectively. These queries are submitted to search the reference document index. The mandatory queries perform conventional search of the Lucene index returning the top 1,000 documents ranked according to equation (1).

CLCGGeoEE5. The query for this run is formulated by topic TD only. The query is submitted to search the footprint document index and reference document index. Search results from the two indexes are combined through linear interpolation equation (2) returning the top 1,000 documents ranked according to equation (2).

CLCGGeoEE10. Two queries are formulated for this run with q_T being the topic TDN content and q_G the geographic-relevant terms extracted from topic TDN. The interpolated similarity score measure in equation (2) is used to rank documents returning the top 1,000.

CLCGGeoEE11. This run is similar to run CLCGGeoEE10 except that geo-relevant terms extracted from topic TDN were expanded by adding geographic references from our geographic resources (see Sect. 2.1).

3 Evaluation and Future Work

3.1 Official GeoCLEF 2006

For runs using interpolated similarity score measure, we set $\lambda_T = \lambda_G = 0.5$. Table 3 shows our official run result. CLCGGeoEE11 is our best run confirming the potential of geographic information (geographic reference expansion) in

improving performance of a conventional IR system. However, several factors might have influenced the performance of our official runs:

- predominance of geographic concept and relation qualifiers such as country, city, southern, west, etc. both in the query and document footprints at the expense of place names, and thereby shifting the query result in the wrong direction propagating irrelevant documents to the top; among the top 20 GeoCLEF 2006 geo-relevant terms were: country, city, street, west, north, south, east, southern, district, town.
- the value of 0.5 assigned to λ_T and λ_G in the interpolated similarity score measure equation (2) might have tilted the result by assigning higher scores to documents retrieved from the reference document collection or vice versa, and thereby propagating irrelevant documents to the top of the rank.

Table 3. Individual Run Performance as measured by Mean Average Precision and R-Precision

	CLCGGeoEE1	CLCGGeoEE2	CLCGGeoEE5	CLCGGeoEE10	CLCGGeoEE11 ^a
MAP	0.1730	0.2163	0.1757	0.1690	0.2194
R-Precision	0.1983	0.2194	0.1777	0.1762	0.2144

^a CLCGGeoEE2 and CLCGGeoEE11 are statistically equivalent because they only differ in the third decimal place.

3.2 Post GeoCLEF 2006

After the release of official GeoCLEF 2006 results a number of experiments were performed to better understand our approach to GIR at CLEF 2006. To confirm our claim to performance influencing factors outlined in Sect. 3.1, we performed geo-query against place-names and assigned values of 0.35 and 0.65 to interpolation factors λ_G and λ_T respectively. The result of the experiment is depicted in Table 4 where column titled Scheme 1 confirms the importance of weighing place-names more than other geographic-relevant terms. Column titled Scheme 2 of Table 4 depicts influence of interpolation factors on search result, and generally the scheme performed better for all runs. Run CLCGGeoEE11b depicted in Table 4 is run CLCGGeoEE11 employing the Porter stemmer and Lucene default stopword list achieving 21.74% improvement over the best score for run CLCGGeoEE11. This result raises the old question of whether GIR is really different from conventional IR [1,5]. We therefore argue that the GIR research community is yet to underpin a clear path to GIR. The community is faced with the basic question of how to understand GIR in order to exploit geographic information within text to improve IR performance for geography constrained user queries. The need for a solid IR system as a basis for building GIR system is clear from the results reported here and by other GeoCLEF 2006 participants [2].

Table 4. Individual run performance as measured by Mean Average Precision and R-Precision

Run	Scheme 1 ^a		Scheme 2 ^b	
	MAP	R-Prec	MAP	R-Prec
CLCGGeoEE1	0.1966	0.2302	0.1966	0.2302
CLCGGeoEE5	0.1976	0.1919	0.2102	0.2109
CLCGGeoEE2	0.2386	0.2312	0.2386	0.2312
CLCGGeoEE10	0.2069	0.1919	0.2388	0.2256
CLCGGeoEE11	0.2397	0.2369	0.2390	0.2307
CLCGGeoEE11b ^c	0.2914	0.2560	0.2918	0.2543

^a geo-query against place-name (see Table 1); $\lambda_G = \lambda_T = 0.5$
^b geo-query against place name (see Table 1); $\lambda_G = 0.35$, $\lambda_T = 0.65$
^c CLCGGeoEE11 employing Porter stemmer and Lucene default stopword list

Geographic Query Expansion. By geographic query expansion (GQE) we mean addition of geographic references obtained from geographic resources to a user query with the goal of improving retrieval performance based on geography. Section 2.1 enumerated geographic resources we used to perform GQE. To assess the impact of GQE, run CLCGGeoEE11b which provided the best performance is chosen and the geographic interpolation factor (GIF) λ_G is varied from 0.0 – 1.0. Equation (2) is used to compute similarity score for vary λ_G . Figure 1 and 2 depicts impact of varying GIF on overall retrieval and per topic performance respectively. TDN-N stands for query formulated by TDN content with no GQE, while TDN-E is formulated by TDN content with GQE applied. Figure 1 shows that GQE stabilizes document retrieval performance. Figure 2 shows topic performance at GIF of 0.0 and 0.8, and again GDN-E stabilizes topic performance across topics. However, topics GC034, GC040 and GC048 seem stable for both GDN-N and GDN-E at GIF of 0.8. A closer look into topics GC040 and GC048 reveals that their main thematic terms *volcano* and *fishing* respectively

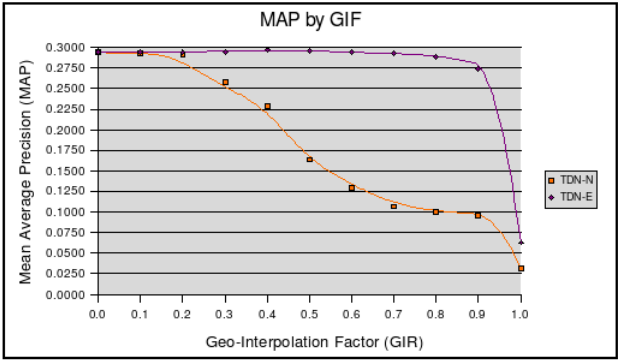


Fig. 1. Document retrieval performance (as measured by MAP) against GIF

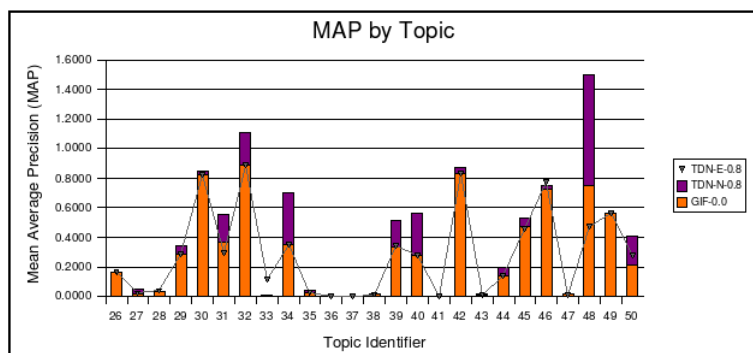


Fig. 2. Topic performance (as measured by MAP) at two GIF (0.0, 0.8)

were tagged as geographic and therefore, the stabilization effect seen for GDN-N. Apart from the stabilization effect provided by GQE, performance of topics GC033 and GC050 was enhanced through GQE, and the performance of topics GC031 and GC048 deteriorated through GQE at GIF of 0.8.

3.3 Future Work

Our future work will focus on investigating new ways to improve the performance of our technique well above standard IR performance along the following lines: 1. a closer look into thematic and geographic terms of topics by applying weighing schemes, 2. optimizing the λ_G value in (2), 3. test other sophisticated geographic similarity score formulae as in [9], 4. exploit GQE stabilization to improve GIR performance, 5. investigate appropriate footprint document indexing strategy, 6. investigate an appropriate document ranking strategy based on geographic information, 7. improve geographic named entity recognition, classification and real world resolution, and 8. test geographic query expansion strategies – blind feedback, addition of place names, expansion through hierarchical information mining, distance based query expansion.

4 Concluding Remarks

We employed a strategy of separately indexing a geographic footprint document in addition to an index of the reference document, and combine the query results of the two indexes through linear interpolation. Our approach yielded an average result as compared to the overall GeoCLEF 2006 result on monolingual English task. GQE was found to stabilize document retrieval performance.

Acknowledgements

This work is supported by NUFFIC within the framework of Netherlands Programme for the Institutional Strengthening of Post-secondary Training Education

and Capacity (NPT) under project titled "Building a sustainable ICT training capacity in the public universities in Uganda".

References

1. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Vivien.: GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santa, D., Rocha, P.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Working Notes for CLEF 2006 Workshop, Alicante, Spain (2006)
3. Gey, F., Petras, V.: Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Larson, R.R.: Cheshire II at GeoCLEF: Fusion and Query Expansion for GIR. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
5. Kornai, A.: MetaCarta at GeoCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
6. Hughes, B.: NICTA i2d2 at GeoCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
7. Buscaldi, D., Rosso, P., Arnal, E.S.: A WordNet-based Query Expansion method for Geographical Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. Gospodnetic, O., Hatcher, E.: Lucene in Action. Manning (2005)
9. Martins, B., Cardoso, N., Chaves, M.S., Andrade, L., Silva, M.J.: The University of Lisbon at GeoCLEF 2006. In: Working Notes for CLEF 2006 Workshop, Alicante, Spain (2006)

GIR with Geographic Query Expansion^{*}

A. Toral, O. Ferrández, E. Noguera, Z. Kozareva, A. Montoyo, and R. Muñoz

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{atoral, ofe, elisa, zkozareva, montoyo, rafael}@dlsi.ua.es

Abstract. For our participation in the second edition of GeoCLEF, we have researched the incorporation of geographic knowledge into Geographic Information Retrieval (GIR). Our system is made up of an IR module (IR-n) and a Geographic Knowledge module (Geonames). The results show that the addition of geographic knowledge has a negative impact on the precision. However, the fact that for some topics the obtained results are better, makes us conclude that the addition of this knowledge could be useful but further research is needed in order to determine how.

1 Introduction

The underlying and basic technology of GIR systems, Information Retrieval (IR), deals with the selection of the most relevant documents from a document collection given a query. Thus, GIR is a specialization of IR which introduces geospatial restrictions to the retrieval task.

Several approaches were followed in order to perform GIR within the first edition of GeoCLEF including using Named Entity Recognition (NER), geographic knowledge resources or NLP tools. There were also proposals based on IR without any treatment of geography.

Three out of the top-4 systems for the English monolingual run were based only on IR (the remaining one [2] used also NER). This may be due to the fact that the systems which tried to use geographic information did not do that in the correct way. Thus, the best results from GeoCLEF 2005, based on IR, may be used as a baseline in order to test the performance of geographic reasoning.

This year we have centered our efforts on studying geographic resources and trying to determine how to use them within GIR. The system developed consists of exploiting geographic resources in order to perform query expansion.

The rest of this paper is organized as follows. The next section presents a description of our system. Section 3 illustrates the experiments and the obtained results. Finally, Section 4 outlines our conclusions and future work proposals.

^{*} This research has been partially funded by the Spanish Government under project CICYT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV06-161.

2 System Description

Our approach is based on IR (see 2.1) with the application of geographic knowledge which is extracted from a structured knowledge resource (see 2.2).

The topics are enriched with geographic information which is obtained by exploiting Geonames¹. A SQL query is generated from the geographic information provided by the topic, and then this query is processed in the Geonames database in order to obtain the related geographic items. Once this information is collected, we apply IR for retrieving the relevant documents.

2.1 IR Module

The IR module we used is called IR-n [3]. It is a Passage Retrieval system (PR). The similarity measure used has been dfr [1].

We have specifically adapted IR-n to incorporate geographic knowledge. In order to do this, we need to take two kinds of restrictions into account :

Required words. These words are marked with ‘#’. Passages which do not contain at least one of these words are not included in the rank list.

Geographical places. In addition, a query expansion has been done using the Geonames database (see 2.2).

As required words, we consider all the nouns in the topic but geographic ones, stop words or other common words appearing in topic definitions. This is, we consider the words that define the main concept of the topic. The objective for doing this is to decrease the amount of noise that the incorporation of big lists of geographic items could introduce.

2.2 Geographic Knowledge Module

Geonames is a geographic database which contains more than 6 million entries for geographical names whereof 2.2 million are cities and villages. Geonames is built from different sources such as nga² and wikipedia³.

Our approach regarding Geonames consisted of building a query to this database for each topic in a methodical way. We extract for each topic the geographic entities and relations, and we enrich the topic with the information that the query with this geographic information returns.

In order to avoid adding noise to the topics and due to the big size of Geonames, we put some restrictions to the extracted data. From the returned entries to the query, we only consider those for which the population is bigger than 10,000 inhabitants and those that belong to a first-order administrative division (ADM1). The geographic queries for all topics are shown in [4].

It should be noted that for some topics (26, 40 and 41), our method to build a query could not be applied, because the topics did not have any geographical restriction considered by Geonames.

¹ www.geonames.org

² <http://gnswww.nga.mil/geonames/GNS/index.jsp>

³ <http://www.wikipedia.org>

3 Experiments and Results

We have evaluated our system for the English and Spanish monolingual tasks. For each task we have carried out three experiments. The first two runs apply basic IR techniques without any geographic reasoning, and we consider these two runs as baselines. The only difference between the first two runs is that the first (*uaTD*) uses only the topic title and description whereas the second (*uaTDN*) uses also the topic narrative.

In the third experiment (*uaTDNGeo*) we also apply IR, but the queries which are passed to this module are previously enriched with geographic information. In the followings paragraphs we show the process that is carried out for this experiment by providing an example for topic 31.

1. Extract from the topic required and geographic words:

```
required words: combat, embargo, effect, fact
geographic entities and relations: Iraq, north-of
```

2. Build the Geonames query

```
SELECT name, alternames FROM geonames
WHERE latitude>33 AND #average latitude of Iraq is 33 N
country_code='IQ' AND ((feature_class='P'
AND population > 10000) OR feature_code='ADM1');
```

3. Assemble a new topic incorporating the extracted geographic knowledge

```
<EN-title>Combats# and embargo# in the northern part of Iraq
</EN-title>
<EN-desc>Documents telling about combats# or embargo# in the
northern part of Iraq</EN-desc>
<EN-narr>Relevant documents are about combats# and effects#
of the 90s embargo# in the northern part of Iraq.
Documents about these #facts happening in other parts of Iraq
are not relevant</EN-narr>
<EN-geonames>Zakho Khurmati Touz [...] </EN-geonames>
```

4. Retrieve the relevant documents using the IR-n system

The addition of geographical information has drastically decremented the precision. For English, the best run (*uaTDN*) obtains 29.85 while the geographic run (*uaTDNGeo*) achieves 12.01 (see Table 1). In the case of Spanish, the best run (*uaTD*) reaches 35.09 and the geographic one (*uaTDNGeo*) 15.25 (see Table 1). Although we implement the model of required words in order to lessen the noise potentially introduced by long lists of geographic items, this seems to be insufficient.

However, for both English and Spanish, the run with geographic information obtains the best results for three topics (see topic by topic results in [4]). Therefore, a more in-depth analysis should be carried out in order to achieve a better understanding on the behaviour of the geographic information incorporation and how it should be done.

Table 1. Overall GeoClef 2006 official results for the Monolingual tasks

Language	Run	AvgP
English	CLEF Average	0.1975
	uaTD	0.2723
	uaTDN	0.2985
	uaTDNGeo	0.1201
Spanish	CLEF Average	0.19096
	uaTD	0.3508
	uaTDN	0.3237
	uaTDNGeo	0.1525

4 Conclusions and Future Work

For our participation in GeoCLEF 2006 we have proposed the expansion of topics with related geographic information. For this purpose we have studied knowledge based geographic resources and we have used Geonames.

The proposal has obtained poor results compared to our simpler model in which we only use IR. This is a paradigmatic example of the state of the art of GIR; it is just the beginning and efforts are needed in order to figure out how to apply geographic knowledge in a way that IR systems could benefit from it.

Therefore, as future work we consider to research into different ways of providing geographic knowledge to basic IR and evaluating the impact of each approach. Thus, our aim is to improve GIR results by applying existing geographic knowledge from structured resources.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic Models of information retrieval based on measuring the divergence from randomness. *Transactions on Information Systems* 20(4), 357–389 (2002)
2. Ferrández, O., Kozareva, Z., Toral, A., Noguera, E., Montoyo, A.: The University of Alicante at GeoCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, Springer, Heidelberg (2005)
3. Llopis, F.: IR-n un Sistema de Recuperación de Información Basado en Pasajes. *Procesamiento del Lenguaje Natural* 30, 127–128 (2003)
4. Toral, A., Ferrández, O., Noguera, E., Kozareva, Z., Montoyo, A., Muñoz, R.: Geographic IR Helped by Structured Geospatial Knowledge Resources. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) *Working Notes at CLEF 2006* (2006)

Monolingual and Bilingual Experiments in GeoCLEF2006

Rocio Guillén

California State University San Marcos, San Marcos, CA 92096, USA
rguillen@csusm.edu

Abstract. This paper presents the results of our experiments in the monolingual English, Spanish and Portuguese tasks and the Bilingual Spanish → English, Spanish → Portuguese, English → Spanish and Portuguese → Spanish tasks. We used the Terrier Information Retrieval Platform to run experiments for both tasks using the Inverse Document Frequency model with Laplace after-effect and normalization 2. Experiments included topics processed automatically as well as topics processed manually. Manual processing of topics was carried out using gazetteers. For the bilingual task we developed a component based on the transfer approach in machine translation. Topics were pre-processed automatically to eliminate stopwords. Then topics in the source language were translated to the target language. Evaluation of results show that the approach worked well in general, but applying more natural language processing techniques would improve retrieval performance.

1 Introduction

Geographic Information Retrieval (GIR) is aimed at the retrieval of geographic data based not only on conceptual keywords, but also on spatial information. Building GIR systems with such capabilities requires research on diverse areas such as information extraction of geographic terms from structured and unstructured data; word sense disambiguation, which is geographically relevant; ontology creation; combination of geographical and contextual relevance; and geographic term translation, among others.

Research efforts on GIR are addressing issues such as access to multilingual documents, techniques for information mining (i.e., extraction, exploration and visualization of geo-referenced information), investigation of spatial representations and ranking methods for different representations, application of machine learning techniques for place name recognition, development of datasets containing annotated geographic entities, among others. [4]. Other researchers are exploring the usage of the World Wide Web as the largest collection of geospatial data.

The purpose of GeoCLEF 2006 is to experiment with and evaluate the performance of GIR systems when topics include geographic locations such as rivers, regions, seas, continents. Two tasks were considered, a monolingual and a bilingual. We participated in the monolingual task in English, Portuguese and Spanish. For the bilingual task we worked with topics in Spanish and documents in

English and Portuguese, and with topics in English and Portuguese and documents in Spanish.

In this paper we describe our experiments in the monolingual task and the bilingual task. Twenty runs were submitted as official runs, thirteen for the monolingual task and seven for the bilingual task. We used the Terrier Information Retrieval (IR) platform to run experiments for both tasks using the Inverse Document Frequency model with Laplace after-effect and normalization 2. Experiments included topics processed automatically as well as topics processed manually. Manual processing of topics was carried out using gazetteers (Alexandria Digital Library, European Parliament and GEOnet Names Server), some of them containing translations in languages other than English, others containing the latitude, longitude and area which allow for semi-automated spatial analysis (proximity analysis). For the bilingual task we developed a component based on the transfer approach in machine translation. Topics were pre-processed automatically to eliminate stopwords. Then topics in the source language were translated to the target language. Terrier was chosen as the retrieval system because it has performed successfully in monolingual information retrieval tasks in CLEF and TREC. Our goal is to have a baseline for further experiments with our component for translating georeferences and spatial analysis as well as applying natural language techniques for disambiguation purposes.

The paper is organized as follows. In Section 2 we present our work in the monolingual task including an overview of Terrier. Section 3 describes our setting and experiments in the bilingual task. We present conclusions and current work in Section 4.

2 Monolingual Task

In this section we first give an overview of Terrier (TERabyte RetRIEveR) an information retrieval (IR) platform used in all the experiments. Then we describe the monolingual experiments for English, Portuguese and Spanish.

Terrier is a platform for the rapid development of large-scale Information Retrieval (IR) systems. It offers a variety of IR models based on the Divergence from Randomness (DFR) framework ([3],[6],[7]). The framework includes more than 50 DFR models for term weighting. These models are derived by measuring the divergence of the actual term distribution from that obtained under a random process ([2]). Terrier provides automatic query expansion with 3 documents and 10 terms as default values; additionally the system allows us to choose a specific query expansion model.

Both indexing and querying of the documents was done with Terrier. The document collections indexed were the LA Times (American) 1994 and the Glasgow Herald (British) 1995 for English, efe94 for Spanish, publico94, publico95, folha94 and folha95 for Portuguese. There were 25 topics for each of the languages tested. Documents and topics in English were processed using the English stopwords list (571 words) built by Salton and Buckley for the experimental SMART IR system [1], and the Porter stemmer. Stopwords lists for Spanish and

Portuguese were also used. No stemmers were applied to the Portuguese and Spanish topics and collections.

We worked with the InL2 term weighting model, which is the Inverse Document Frequency model with Laplace after-effect and normalization 2. Our interpretation of GeoCLEF's tasks was that they were not exactly classic ad-hoc tasks, hence we decided to use a model for early precision. We experimented with other models and found out that this model generated the best results when analyzing the list of documents retrieved.

The risk of accepting a term is inversely related to its term frequency in the document with respect to the elite set, a set in which the term occurs to a relatively greater extent than in the rest of the documents. The more the term occurs in the elite set, the less the term frequency is due to randomness. Hence the probability of the risk of a term not being informative is smaller. The Laplace model is utilized to compute the information gain with a term within a document. Term frequencies are calculated with respect to the standard document length using a formula referred to as normalization 2 shown below.

$$tfn = tf \cdot \log\left(1 + c \frac{sl}{dl}\right)$$

tf is the term frequency, sl is the standard document length, and dl is the document length, c is a parameter. We used $c = 1.5$ for short queries, which is the default value, $c = 3.0$ for short queries with automatic query expansion and $c = 5.0$ for long queries. Short queries in our context are those which use only the topic title and topic description; long queries are those which use the topic title, topic description and topic narrative. We used these values based on the results generated by the experiments on tuning for BM25 and DFR models done by He and Ounis [5]. They carried out experiments for TREC (Text REtrieval Conference) with three types of queries depending on the different fields included in the topics given. Queries were defined as follows: 1) short queries are those where the title and the description fields are used; and 2) long queries are those where title, description and narrative are used.

2.1 Experimental Results

We submitted 5 runs for English (one of them was repeated), 4 runs for Portuguese and 5 runs for Spanish. For some of the runs we used the automatic query expansion capability of terrier with the default values of 3 documents and 10 terms. A major problem we detected after submitting our results was that we did not include the Spanish newspaper collection for the year 95 (EFE 95) for indexing and retrieval purposes. Therefore, the results of our experiments with Spanish for the monolingual and bilingual tasks were affected in terms of recall and precision. Results for the monolingual task in English, Portuguese and Spanish are shown in Table 1, Table 2 and Table 3, respectively.

Official evaluation of results described by the organizers in [11] for the English Monolingual task (title and description only) show that our best run was SMGeoEN4 not pooled. It ranked third among the top five runs.

Table 1. English Monolingual Retrieval Performance

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoEN1	title, description	automatic	yes	26.37	28.57
SMGeoEN3	title, description, narrative	automatic	yes	28.57	33.66
SMGeoEN4	title, description	automatic	no	26.37	28.57
SMGeoEN5	title, description, narrative	automatic	no	23.77	25.81

Table 2. Portuguese Monolingual Retrieval Performance

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoPT1	title, description, narrative	automatic	yes	10.98	13.91
SMGeoPT2	title, description	automatic	yes	13.44	15.02
SMGeoPT3	title, description, narrative	automatic	no	10.98	13.91
SMGeoPT4	title, description	automatic	no	10.63	13.57

Official evaluation of results described by the organizers in [11] for the Portuguese Monolingual task (title and description only) show that our best run was SMGeoPT2 pooled.

Table 3. Spanish Monolingual Retrieval Performance

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoES1	title, description, narrative	automatic	yes	14.71	20.44
SMGeoES2	title, description	automatic	yes	14.71	20.44
SMGeoES3	title, description	manual	yes	14.71	20.44
SMGeoES4	title, description	automatic	no	13.78	18.63
SMGeoES5	title, description, narrative	automatic	no	14.71	20.44

Official evaluation of results described by the organizers in [11] for the Spanish Monolingual task (title and description only) show that our best run was SMGeoES1 pooled.

We have run experiments with the full Spanish collection. Additional experiments using the Ponte-Croft language modeling feature of Terrier were done. The basic idea behind this approach is to estimate a language model for each document, and rank documents by the likelihood of the query according to the language model [12]. The weight of a word in a document is given by a probability. For query $Q = q_1 q_2 \dots q_n$ and document $D = d_1 d_2 \dots d_m$, this probability is denoted by $p(q | d)$. Documents are ranked estimating the posteriori probability $p(d | q)$; applying Bayes' rule $p(d | q) = p(q | d)p(d)$, where $p(d)$ represents the prior belief that d is relevant to any query and $p(q | d)$ is the query likelihood

Table 4. Full Spanish Collection Monolingual Retrieval Performance

Run Id	Topic Fields	Query Construction	Query Expansion	MAP	Recall Prec.
NGeo1	title, description	automatic	no	22.59	24.71
NGeo2	title, description	automatic	yes	26.48	27.61
NGeo3	title, description	manual	no	17.14	18.87
NGeo4	title, description, narrative	automatic	no	20.79	23.31
NGeo5	title, description, narrative	automatic	yes	20.79	23.31
NGeo6	title, description, narrative	manual	no	19.91	21.32
NGeo7	title, description, narrative	manual	yes	23.30	22.88
NGeo8	title, description, narrative	manual	no	17.43	20.39

given the document. The Ponte-Croft model assumes that $p(d)$ is uniform, and so does not affect document ranking.

Run NGeo1 was processed with the title and description tags for retrieval, queries were automatically constructed and no query expansion was done. Run NGeo2 was processed with the title and description tags, queries were automatically constructed and query expansion was performed with the default values (top 3 documents and 10 terms). For NGeo3 we constructed the queries manually with the title and description tags, no query expansion was done.

Five runs with the title, description and narrative were processed. Queries for run NGeo4 were built automatically and no query expansion was done. For run NGeo5 we did query expansion, queries were built automatically. Manual queries were built for runs NGeo6, NGeo7 and NGeo8 and were processed with no query expansion, with query expansion, and with the Ponte-Croft model, respectively. These queries used the Spanish Toponymy from the European Parliament [9], and the Names files of countries and territories from the GEOnet Names Server (GNS) [10]. Evaluation of the new runs was done with *trec_eval* [13] and the results are presented in Table 4.

Term weighting in the DFR models is calculated differently from the Ponte-Croft language model, which may explain the difference in the retrieval performance of runs NGeo6, NGeo7 and NGeo8.

Query expansion improved results for runs processing title and description as well as for runs including the narrative. Manually built queries did not perform as well without query expansion and without the narrative. This may be because we introduced many terms that were not in the documents. For instance, we manually built the tag title for Topic 27 as follows:

< ES – title > Maguncia, Renania, Renania-Palatinado, Kaiserlautern, Bundesbank, Mannheim Mannheim Wiesbaden Francfort Frankfurt < /ES – title >,

and the original contents of the title tag for Topic 027, automatically built, were as follows:

< ES – title > Ciudades a menos de 100 kilometros de Francfort < /ES – title >.

3 Bilingual Task

For the bilingual task we worked with Spanish topics and English and Portuguese documents, and English and Portuguese topics and Spanish documents. We built a component, independent of Terrier, based on the transfer approach in machine translation to translate topics from the source language to the target language using mapping rules. All of the information in the topics within the title, description and narrative was translated. Topics in English, Spanish, and Portuguese were preprocessed by removing diacritic marks and using stopwords lists. Diacritic marks were also removed from the stopwords lists and duplicates were eliminated. Plural stemming was then applied.

Automatic query construction was carried out with the aid of the German Alexandria Digital Library gazetteer [8], the Spanish Toponymy from the European Parliament [9], and the Names files of countries and territories from the GEOnet Names Server (GNS) [10]. The German gazetteer was particularly helpful because it included information such as latitude, longitude and area. Thus, English Topic 027 with narrative “Relevant documents discuss cities within 100 kilometers of Frankfurt am Main Germany, latitude 50.11222, longitude 8.68194...” lend itself to spatial analysis using a distance measure to find out the cities within 100 kilometers of Frankfurt.

3.1 Experimental Results

Seven runs were submitted as official runs for the GeoCLEF2006 bilingual task. In Table 5 we report the results for X-Spanish ($X=\{\text{English, Portuguese}\}$) and in Table 6 the results for Spanish-X ($X=\{\text{English, Portuguese}\}$).

Table 5. X-Spanish Bilingual Retrieval Performance ($X = \{\text{English, Portuguese}\}$)

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoENES1	title, description	automatic	no	12.82	16.89
SMGeoPTES2	title, description	automatic	no	10.89	14.67
SMGeoPTES3	title, description, narrative	automatic	no	11.50	15.27

Official evaluation of results described by the organizers in [11] for the Bilingual tasks (title and description only) show the following:

- SMGeoESEN2 pooled ranked second among the best entries for Bilingual English.
- SMGeoESPT2 pooled ranked first among the best entries for Bilingual Portuguese.
- SMGeoENES1 pooled ranked second for Bilingual Spanish.

Table 6. Spanish-X Bilingual Retrieval Performance ($X = \{\text{English, Portuguese}\}$)

Run Id	Topic Fields	Query Construction	Query Expansion	Avg Prec.	Recall Prec.
SMGeoESEN1	title, description	automatic	no	12.82	16.89
SMGeoESEN2	title, description, narrative	automatic	no	12.82	16.89
SMGeoESPT1	title, description, narrative	automatic	no	10.89	14.67
SMGeoESPT2	title, description	automatic	no	11.50	15.27

4 Conclusions

In this paper we presented work on monolingual and bilingual geographical information retrieval. We used Terrier to run our experiments, and an independent translation component built to map source language (English, Portuguese or Spanish) topics into target language (English, Portuguese or Spanish) topics. Official evaluation of the results show that the approaches used in the monolingual and bilingual tasks perform well in general, but more natural language processing techniques may improve retrieval performance. The results for Spanish were good even when we did not include the full collection of documents. We ran new experiments with the complete Spanish collection and evaluation of the results show that the DFR model performed better than the Ponte-Croft model. In addition, query expansion improved the retrieval performance for short (title and description) and for long queries (title, description and narrative).

References

1. <http://ftp.cs.cornell.edu/pub/smart/>
2. Lioma, C., He, B., Plachouras, V., Ounis, I.: The University of Glasgow at CLEF2004; French monolingual information retrieval with Terrier. In: Working notes of the CLEF 2004 Workshop, Bath, UK (2004)
3. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, Springer, Heidelberg (2005), <http://ir.dcs.ga.ac.uk/terrier/>
4. Purves, R., Jones, C.: SIGIR2004: Workshop on Geographic Information Retrieval, Sheffield, UK (2004)
5. He, B., Ounis, I.: A study of parameter tuning for the frequency normalization. In: Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA
6. Amati, G., van Rijsbergen, C.J.: Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems* 20(4), 357–389
7. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006) (2006)

8. Alexandria Digital Library Gazetteer. 1999- Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents. (ADL Gazetteer Development page with links to various clients and protocols that access the ADL Gazetteer), <http://www.alexandria.ucsb.edu/gazetteer>
9. European Parliament. Tools for the External Translator, http://www.europarl.europa.eu/transl_es/plataform/pagina/toponim/toponimo.htm
10. <http://earth-info.nga.mil/gns/html/index.html>
11. Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P.: GeoCLEF 2006: the CLEF2006 Cross-Language Geographic Information Retrieval Track Overview. In: Working notes of the CLEF 2006 Workshop, Alicante, Spain (2006)
12. Zhai, C., Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems* 22(2) (2004)
13. Text REtrieval Conference, <http://trec.nist.gov>

Experiments on the Exclusion of Metonymic Location Names from GIR

Johannes Leveling and Dirk Veiel

FernUniversität in Hagen (University of Hagen)
Intelligent Information and Communication Systems (IICS)
58084 Hagen, Germany
`firstname.lastname@fernuni-hagen.de`

Abstract. For the GeoCLEF task of the CLEF campaign 2006, we investigate identifying literal (geographic) and metonymic senses of location names (the location name refers to another, related entity) and indexing them differently. In document preprocessing, location name senses are identified with a classifier relying on shallow features only. Different senses are stored in corresponding document fields, i. e. LOC (all senses), LOCLIT (literal senses), and LOCMET (metonymic senses). The classifier was trained on manually annotated data from German CoNLL-2003 data and from a subset of the GeoCLEF newspaper corpus.

The setup of our GIR (geographic information retrieval) system is a variant of our setup for GeoCLEF 2005. Results of the retrieval experiments indicate that excluding metonymic senses of location names (short: metonymic location names) improves mean average precision (MAP). Furthermore, using topic narratives decreases MAP, and query expansion with meronyms improves the performance of GIR in our experiments.

1 Introduction

Geographic information retrieval (GIR) applications have to perform tasks such as the identification and disambiguation of location names (see [1] for a general overview over tasks in GIR applications). Differentiating between literal (geographic) and metonymic location names should be a part of the disambiguation process. Metonymy is typically defined as a figure of speech in which a speaker uses “*one entity to refer to another that is related to it*” [2]. For example, in the sentence *Germany and Paris met in Rome to decide about import taxes*, *Rome* is used in its geographic sense while *Germany* and *Paris* are used in a metonymic sense and refer to a group of officials (see [3] for an overview over regular metonymy for location names). A text containing this sentence would not be relevant to the query *Find documents about taxes in Germany*. For our participation at the GeoCLEF task in 2006, we investigated excluding metonymic location names to avoid retrieving documents containing metonymic names.

We employ a classifier for differentiating between literal and metonymic senses of location names [4]. Different senses correspond to different indexes for location names: LOC (containing all senses), LOCLIT (literal senses), and LOCMET

(metonymic senses). The experiments focus on metonym identification in *documents*, because the GeoCLEF query topics did not contain any metonymic location names in a topic title, description, or narrative.

2 System Description

Our GIR system is based on the system developed for GeoCLEF 2005 [5]. A syntactic-semantic parser [6] analyzes the query topics and the GeoCLEF corpus of newspaper and newswire articles. From its parse results, base forms (lemmata) and compound constituents are extracted as search terms or index terms.

The major change made to the GeoCLEF 2005 system lies in treating metonymic location names in documents in a different way. For the identification of metonymic location names, we utilized a classifier trained on a manually annotated subset of the German CoNLL-2003 Shared Task corpus for Language-Independent Named Entity Recognition [7] and a subset of the GeoCLEF newspaper corpus. The metonymy classifier [4] is based on shallow features (i. e. not based on syntactic or semantic features) only, including part-of-speech tags for closed word classes from a list lookup, position of words in a sentence, word length, and base forms of verbs. It achieved a performance of 81.7% F_1 -measure in differentiating between literal and metonymic location names. In analyzing the annotated CoNLL data (1216 instances), we found that 16.95% of all location names were used metonymically, and 7.73% referenced both a literal and a metonymic sense at the same time (see [4] for a more detailed description of the metonymy classification). These numbers indicate an upper limit of about 25% performance increase for methods exploiting metonymy information.

After preprocessing, the documents are structured with the following fields: document ID (DOCID), text of the document (TEXT), location names from the text (LOC), location names in their literal sense (LOCLIT), and location names in their metonymic sense (LOCMET). The representations of all documents in the CLEF newspaper collection were indexed with the Zebra database management system [8], which supports a standard relevance ranking (*tf-idf* IR model).

3 Description of the Submitted Runs and Results

Two methods were employed to obtain an IR query from a topic title, description, and narrative: 1.) The parser performs a syntactic-semantic analysis of the query text and the resulting semantic net is transformed into a database independent query representation (DIQR, see [9]). 2.) Topical search terms extracted from the query text are combined with location names (identified by a name lookup) to obtain the DIQR. In both cases, a DIQR query consists of a Boolean combination of topical search terms (or descriptors) and location names.

In our baseline experiment (FUHddGNNNTD), the standard IR model (*tf-idf*) was utilized without additions or modifications. In experiments not accessing a separate name index, location names were searched within the index for the TEXT field; in experiments using the separate index (e. g., FUHddGYYYTDN)

Table 1. Mean average precision (MAP) and relevant and retrieved documents (rel_ret) for monolingual German GeoCLEF experiments. 785 documents were assessed as relevant for the 25 queries in the GeoCLEF 2006 test set.

Run Identifier	Language		Parameters						Results	
	Source	Target	LI	LA	QEX	MET	QF		MAP	rel_ret
FUHddGNNNTD	DE	DE	N	N	N	N	TD		0.1694	439
FUHddGYYYTD	DE	DE	Y	Y	Y	N	TD		0.2229	449
<i>FUHddGYYNTD</i>	DE	DE	Y	Y	N	N	TD		0.1865	456
FUHddGNNNTDN	DE	DE	N	N	N	N	TDN		0.1223	426
FUHddGYYYTDN	DE	DE	Y	Y	Y	N	TDN		0.2141	462
FUHddGYYYMTDN	DE	DE	Y	Y	Y	Y	TDN		0.1999	442

location names were looked for in the index for the LOC field. For experiments excluding metonymic senses, the index for location names corresponds to the field LOCLIT, i.e. only location names in their literal sense were searched for.

The settings for a range of parameters were varied for the submitted runs. A separate location index corresponding to the LOC or LOCLIT field (parameter LI=Y) or the TEXT field may be accessed (LI=N). The syntactic-semantic parser (see [6]) is applied and the resulting semantic net is transformed into a database query (LA=Y, see [9]) or bag-of-words IR is performed (LA=N). Semantically related terms and meronyms are employed for query expansion (QEX=Y, [5]) or no query expansion takes place (QEX=N). Metonymic location name senses are excluded (MET=Y, [4]) or not (MET=N). The query topic fields title and description are accessed (QF=TD) or topic title, description, and narrative are utilized (QF=TDN). Table 1 shows the different parameter settings and results for our monolingual German GeoCLEF runs. The results include mean average precision (MAP) and the number of relevant and retrieved documents (rel_ret) in the set of 25,000 documents per experiment.

4 Analysis of the Results

Using a separate index for location names and removing metonymic location names from indexing leads to a better performance in almost all our experiments. Significance tests have been omitted, because the number of queries (25) is still too small to obtain meaningful results.¹

In addition to indexing metonymic location names, we experimented with query expansion and including the topic narrative to obtain additional search terms. As we observed from results of the GeoCLEF 2005 experiments [5], query expansion with meronyms (semi-automatically collected part-of relations from gazetteers) leads to significantly better precision for most monolingual German runs.

¹ System performance for queries from the GeoCLEF track in 2005 was too low to include them for testing.

One hypothesis to be tested was that the additional information in topic narratives (runs with QEX=TDN instead of TD) would improve results. We can not confirm this assumption because with our setup, MAP and relevant and retrieved documents are almost always lower for runs using the topic narrative than for runs with topic title and description only.

5 Conclusion and Future Work

We investigated whether exclusion of metonymic location names from indexing would increase precision in GIR experiments. A significant increase in precision for experiments excluding metonymic location names was expected. This assumption holds for all experiments conducted (except one), confirming results of earlier experiments [4]. Summarizing, the results look promising, but a larger set of geographic questions is needed to perform meaningful significance testing.

Future work will investigate if instead of completely removing all metonymic location names from the database index, the corresponding terms should be processed differently (with a lesser weight or with a different sense).

Additional information from the topic narratives did not improve precision or recall in our experiments, although one might have expected a similar effect as for query expansion with meronyms.

References

1. Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M.J., Weibel, R.: Spatial information retrieval and geographical ontologies – an overview of the SPIRIT project. In: Proceedings of SIGIR 2002, pp. 387–388 (2002)
2. Lakoff, G., Johnson, M.: Metaphors we live by. Chicago University Press, Chicago (1980)
3. Markert, K., Nissim, M.: Towards a corpus for annotated metonymies: the case of location names. In: Proceedings of the 3rd international Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain (2002)
4. Leveling, J., Hartrumpf, S.: On metonymy recognition for GIR. In: Proceedings of the 3rd ACM workshop on GIR (2006)
5. Leveling, J., Hartrumpf, S., Veiel, D.: Using semantic networks for geographic information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 977–986. Springer, Heidelberg (2006)
6. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück, Germany (2003)
7. Sang, E.T.K., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: Language independent named entity recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of CoNLL-2003, Edmonton, Canada, pp. 142–147 (2003)
8. Hammer, S., Dickmeiss, A., Levanto, H., Taylor, M.: Zebra – user’s guide and reference. Manual, IndexData, Copenhagen, Denmark (2005)
9. Leveling, J.: Formale Interpretation von Nutzeranfragen für natürlichsprachliche Interfaces zu Informationsangeboten im Internet. Dissertation, Fachbereich Informatik, FernUniversität in Hagen (2006)

The University of New South Wales at GeoCLEF 2006

You-Heng Hu and Linlin Ge

School of Surveying and Spatial Information Systems
University of New South Wales, Sydney, Australia
you-heng.hu@student.unsw.edu.au, l.ge@unsw.edu.au

Abstract. This paper describes our participation in the GeoCLEF monolingual English task of the Cross Language Evaluation Forum 2006. The main objective of this study is to evaluate the retrieve performance of our geographic information retrieval system. The system consists of four modules: the geographic knowledge base that provides information about important geographic entities around the world and relationships between them; the indexing module that creates and maintains textual and geographic indices for document collections; the document retrieval module that uses the Boolean model to retrieve documents that meet both textual and geographic criteria; and the ranking module that ranks retrieved results based on ranking functions learned using Genetic Programming. Experiments results show that the geographic knowledge base, the indexing module and the retrieval module are useful for geographic information retrieval tasks, but the proposed ranking function learning method doesn't work well.

Keywords: Geographic Information Retrieval, geographic knowledge base, geo-textual indexing, Genetic Programming.

1 Introduction

The GeoCLEF track of the Cross Language Evaluation Forum (CLEF) aims to provide a standard test-bed for retrieval performance evaluation of Geographic Information Retrieval (GIR) systems using search tasks involving both geographic and multilingual aspects [2].

Five key challenges have been identified in building a GIR system for GeoCLEF 2006 tasks:

Firstly, a comprehensive geographic knowledge base that provides not only flat gazetteer lists but also relationships between geographic entities must be built, as it is essential for geographic references extraction and grounding during all GIR query parsing and processing procedures.

Secondly, comparing with GeoCLEF 2005, GeoCLEF 2006 no longer provides explicit expressions for geographic criteria. Topics must be geo-parsed to identify and extract geographic references that are embedded in the title, description and narrative tags. In addition, new geographic relationships are introduced in GeoCLEF 2006, such as geographic distances (e.g. within 100km of Frankfurt) and complex geographic expressions (e.g. Northern Germany).

Thirdly, an efficient geo-textual indexing scheme must be built for efficient document accessing and searching. Although the computation cost is not considered in the system evaluation, the indexing scheme is necessary for a practical retrieval system where large numbers of documents are involved.

Fourthly, the relevance of a document to a given topic in GIR must be determined not only by thematic similarity, but also by geographic associations between them.

And lastly, a uniform ranking function that takes into account both textual and geographic similarity measures must be specified to calculate a numerical ranking score for each retrieved document.

This is our first participation in the CLEF tasks. The main objective of this study is to evaluate the performance of the GIR system developed at the School of Surveying and Spatial Information Systems at the University of New South Wales, Australia.

The remainder of the paper is organised as follows: Section 2 describes the design and implement of our GIR system. Section 3 presents our runs carried out for the monolingual English task. Section 4 discusses the obtained results. And finally, Section 5 concludes the paper and gives future work directions.

2 Approaches for GeoCLEF 2006

This section describes the specific approaches for our participation in the GeoCLEF 2006 monolingual English task.

2.1 System Overview

Our proposed methodology in the development of a GIR system includes a geographic knowledge base for representation and organisation of geographic data and knowledge, an integrated geo-textual indexing scheme for document searching, a Boolean model [6] for document retrieval and a ranking function learning algorithm based on Genetic Programming (GP) [1, 3].

Figure 1 shows the overview architecture of our system. The JAVA programming language is used to implement the whole system; the MySQL database is used as the

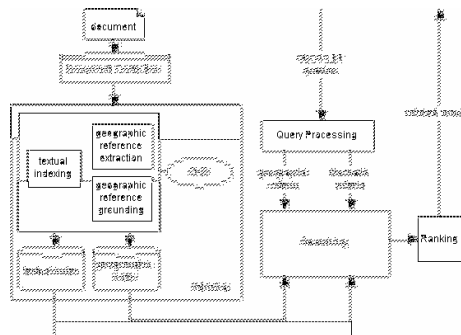


Fig. 1. System architecture of the GIR system used in the GeoCLEF 2006

backend database; the Apache Lucene search engine (c.f. <http://lucene.apache.org>) is used for textual indexing and searching, and the Alias-I LingPipe system (c.f. <http://www.alias-i.com/lingpipe>) is used for the Named Entity Recognition (NER) task.

2.2 Geographic Knowledge Base

Data in our geographic knowledge base is collected from various public sources and compiled into the MySQL database. The statistics of our geographic knowledge base are given in Table 1.

Table 1. Statistics for the geographic knowledge base

<i>Description</i>	<i>Statistic</i>
Number of distinct geographic entities/names	7817/8612
- Number of countries/names	266/502
- Number of administrative divisions/names	3124/3358
- Number of cities/names	3215/3456
- Number of oceans, seas, gulfs, rivers/names	849/921
- Number of regions/names	363/375
Average names per entity	1.10
Number of relationships	9287
- Number of part-of relationships	8203
- Number of adjacency relationships	1084
Number of entities that have only one name	7266 (92.95%)
Number of entities without any relationship	69 (0.88%)
Number of entities without any part-of relationship	123 (1.57%)
Number of entities without any adjacency relationship	6828 (87.35%)

2.3 Textual-Geo Indexing

Our system creates and maintains the textual index and the geographic index separately, and links them using the document identifications.

The textual index is built using Lucene with its build-in support for stop words removing [5] and the Porter stemming algorithm [4].

The geographic index is built as a procedure of three steps. The first step performs a simple string matching against all documents in the collections utilising the place name list derived from our geographic knowledge base. The second step performs a NER process to tag three types of named entities: PERSON, LOCATION and ORGANISATION. The final step matches result sets from the two previous steps using following rules: (1) for each string that found in the first step, it is eliminated if it is tagged as a non-location entity in the second step, otherwise it is added to the geographic index; (2) for each place name in the stop word list of the first step, it is added to the geographic index if it is tagged as a location entity in the second step.

2.4 Document Retrieval

The retrieval of relevant documents in our system is a four-phase procedure that involves query parsing, textual searching, geographic searching and Boolean intersection.

The GeoCLEF query topics are in general modelled as $Q = (\text{textual criteria}, \text{geographic criteria})$ in our system. However, the query parser is configured in an ad hoc fashion for the GeoCLEF 2006 tasks at hand. Given a topic, the parser performs the following steps: (1) Removes guidance information, such as “Documents about” and “Relevant documents describe”. Description about irrelevant documents is removed as well. (2) Extracts geographic criteria using string matching with names and types data obtained from the geographic knowledge base. The discovered geographic entities, geographic relationships and geographic concepts are added to the geographic criteria. Then geographic related words are removed from the query topic. (3) Removes stop words. (4) Expands all-capitalised abbreviations (e.g. *ETA* in GC049) using the WordNet APIs. Then, the result text is treated as textual keywords.

After query topics are parsed, the Lucene search engine is used to retrieve all documents that contain the textual keywords, and the geographic index is used to retrieve all documents that meet the geographic criteria.

Having retrieved the two results sets, the final step intersects them using the Boolean AND operator, only documents that appear in both result sets are considered as relevant documents.

2.5 Document Ranking

A Genetic Programming-based algorithm is developed in our system to learn ranking functions. Our GP algorithm consists of two elements: (1) a set of terminals and functions that can be used as logic unit of a ranking function; and (2) a fitness function that is used to evaluate the performance of each candidate ranking function.

Terminals reflect logical views of documents and user queries. Table 2 lists terminals used in our system. Functions used in our experiments include addition (+), subtraction (-), multiplication ($\times$), division (/) and natural logarithm (log).

Table 2. Terminals used in the ranking function learning process

<i>Name</i>	<i>Description</i>
<i>DOC_COUNT</i>	number of documents in the collection
<i>DOC_LENGTH</i>	length of the document
<i>LUCENE_SCORE</i>	Lucene ranking score of the document
<i>GEO_NAME_NUM</i>	how many geographic names in the document
<i>GEO_NAME_COUNT</i>	total number of geographic names of all geographic entities discovered from the document
<i>GEO_ENTITY_COUNT</i>	how many entities that have the geographic name
<i>GEO_RELATED_COUNT</i>	how many entities that have the geographic name and related to the query
<i>GEO_NAME_DOC_COUNT</i>	number of documents that have the geographic name
<i>GEO_COUNT</i>	how many times of the geographic name appears in the document
<i>NAME_COUNT</i>	number of geographic names in the geographic knowledge base
<i>ENTITY_COUNT</i>	number of entities in the geographic knowledge base

Three fitness functions are used in our system:

F_P5. This fitness function returns the arithmetic mean of the precision values at 50% recall for all queries as results.

$$F_{P5} = \frac{1}{Q} \times \sum_{i=1}^Q P_{i,5} \quad (1)$$

where Q is the total number of queries, $P_{i,5}$ is the precision value at fifth recall level of the 11 standard recall level (i.e. 50% recall) of the i th query.

F_{MAP} . This fitness function utilises the idea of average precision at seen relevant documents.

$$F_{MAP} = \frac{1}{Q} \times \sum_i AP_i \quad (2)$$

$$AP_i = \frac{1}{R_i} \times \sum_{j=1}^{R_i} r(d_j) \times \frac{\sum_{k=1}^j r(d_k)}{j} \quad (3)$$

where R_i is the total number of relevant documents for the i th query, $r(d_j)$ is a function that returns 1 if the document d_j is relevant for i th query and returns 0 otherwise.

F_{WP} . This fitness function utilises the weighted sum of precision values on the 11 standard recall levels.

$$F_{WP} = \frac{1}{Q} \times \sum_{i=1}^Q WP_i \quad (4)$$

$$WP_i = \sum_{j=0}^{10} \frac{1}{(i+1)^m} P_{i,j} \quad (5)$$

where $P_{i,j}$ is the precision value at the j th recall level of the 11 standard recall levels for the i th query, m is a positive scaling factor determined from experiments.

3 Experiments

Following five runs were submitted for GeoCLEF 2006 monolingual English tasks:

unswTitleBase: This run used the title and description tags of the topics for query parsing and searching. After relevant documents were retrieved, the Lucene ranking scores were used to rank results.

unswNarrBase: This run used the title, description and narrative tags of the topics for query parsing and searching. After relevant documents were retrieved, the Lucene ranking scores were used to rank results.

The above two runs were mandatory, and they were used as our baseline methods.

unswTitleF46: This run used the title and description tags of the topics for query parsing and searching. After relevant documents were retrieved, the ranking function given below was used to rank results. This ranking function was discovered using fitness function F_{WP} with $m = 6$.

$$LUCENE\_SCORE * LUCENE\_SCORE * LUCENE\_SCORE / GEO\_NAME\_COUNT$$

unswNarrF41: This run used the title, description and narrative tags of the topics for query parsing and searching. After relevant documents were retrieved, the ranking

function given below was used to rank results. This ranking function was discovered using fitness function F_{WP} with $m = 1$.

$$LUCENE\_SCORE * LUCENE\_SCORE * LUCENE\_SCORE * GEO\_RELATED\_COUNT / DOC\_LENGTH$$

unswNarrMap: This run used the title, description and narrative tags of the topics for query parsing and searching. After relevant documents were retrieved, the ranking function given below was used to rank results. This ranking function was discovered using fitness function F_{MAP} .

$$GEO\_RELATED\_COUNT * LUCENE\_SCORE / DOC\_COUNT / DOC\_COUNT$$

The ranking functions used in the above three runs were learned using our GP learning algorithm. The GeoCLEF 2005 topics and relevance judgments are used as training data.

4 Results

Table 3 summarises the results of our runs using evaluation metrics include Average Precision, R-Precision and the increment over the mean average precision (i.e. 19.75%) obtained from all submitted runs. The precision average values for individual queries are shown in Table 4.

Table 3. GeoCLEF 2006 monolingual English tasks results

<i>Run</i>	<i>AvgP. (%)</i>	<i>R-Precision (%)</i>	<i>Δ AvgP. Diff over GeoCLEF Avg P. (%)</i>
unswTitleBase	26.22	28.21	+32.75
unswNarrBase	27.58	25.88	+39.64
unswTitleF46	22.15	26.87	+12.15
unswNarrF41	4.01	4.06	-79.70
unswNarrMap	4.00	4.06	-79.75

Several observations are made from the obtained results: firstly the geographic knowledge base and the retrieve model used in our system showed their potential usefulness in GIR as we can see from the higher average precision values of unswTitleBase (26.22%) and unswNarrBase (27.58%), which archived a 32.75% and a 39.64% improvement comparing to the mean average precision of all submitted runs.

Secondly, the ranking function learning algorithm used in our system doesn’t work well for GeoCLEF tasks, particular for those runs (i.e. unswNarrF41 and unswNarrMap) that utilise narrative information of the queries. We suppose such behaviour is due to a strong over-training effect. However, the unswTitleF46 run performed better than the two base line runs in a small set of topics (i.e. GC027, GC034, GC044 and GC049).

Thirdly, it is not immediately obvious that the narrative information should be included in the query processing. The unswTitleBase run achieved the same performance as the unswNarrBaseline run in 10 topics (i.e. GC026, GC027, GC030,

GC036, GC037, GC040, GC041, GC046, GC049 and GC050), and it even achieved better results in 6 topics (i.e. GC028, GC029, GC033, GC034, GC039 and GC044).

Lastly, it is interesting to see that our system didn't retrieve any relevant document for topic GC036 and GC041. It is not surprised for GC036, as no document was identified as relevant in the assessment result. While for GC041, which talks about "Shipwrecks in the Atlantic Ocean", the keyword "shipwreck" doesn't appear in any of the four relevant documents.

Table 4. Precision average values (%) for individual queries

<i>Topic</i>	<i>unswTitleBase</i>	<i>unswNarrBase</i>	<i>unswTitleF46</i>	<i>unswNarrF41</i>	<i>unswNarrMap</i>
GC026	30.94	30.94	15.04	0.58	0.56
GC027	10.26	10.26	12.32	10.26	10.26
GC028	7.79	3.35	5.09	0.36	0.31
GC029	24.50	4.55	16.33	0.53	0.53
GC030	77.22	77.22	61.69	6.55	6.55
GC031	4.75	5.37	5.09	3.31	3.31
GC032	73.34	93.84	53.54	5.71	5.71
GC033	46.88	38.88	44.77	33.71	33.71
GC034	21.43	2.30	38.46	0.14	0.13
GC035	32.79	43.80	28.06	3.19	3.11
GC036	0.00	0.00	0.00	0.00	0.00
GC037	21.38	21.38	13.17	0.81	0.81
GC038	6.25	14.29	0.12	0.12	0.12
GC039	46.96	45.42	34.07	3.50	3.50
GC040	15.86	15.86	13.65	0.34	0.30
GC041	0.00	0.00	0.00	0.00	0.00
GC042	10.10	36.67	1.04	0.33	0.33
GC043	6.75	16.50	4.33	0.55	0.54
GC044	21.34	17.23	13.80	4.78	4.78
GC045	1.85	3.96	2.38	1.42	1.42
GC046	66.67	66.67	66.67	3.90	3.90
GC047	8.88	11.41	9.80	1.02	0.98
GC048	58.52	68.54	51.55	8.06	8.06
GC049	50.00	50.00	50.00	0.09	0.09
GC050	11.06	11.06	12.73	11.06	11.06

5 Conclusions

This paper proposed the GIR system that has been developed for our participation in the GeoCLEF 2006 monolingual English task. The key components of the system, including a geographic knowledge base, an integrated geo-textual indexing scheme, a Boolean retrieval model and a Genetic Programming-based ranking function discovery algorithm, are described in detail. The results shows that the geographic knowledge base, the indexing module and the retrieval model are useful for geographic information retrieval tasks, but the proposed ranking function learning method doesn't work well.

Clearly there is much work to be done in order to fully understand the implications of the experiment results. The future research directions that we plan to pursue

include: (1) the establishing of a unified GIR retrieval model that is capable to combine textual and geographic representation and ranking of documents in a suitable framework: (2) the utilising of parallel computation techniques to improve the system computation performance and (3) the extending of our geographic knowledge base by adding more feature types, such as population number and economic importance, which may affect relevance judgment and ranking.

References

1. Fan, W.P., Gordon, M.D.P., Pathak, P.P.: Genetic Programming-Based Discovery of Ranking Functions for Effective Web Search. *Journal of Management Information Systems* 21(4), 37–56 (2005)
2. Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Nunzio, G.M.D., Ferro, N.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview'. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) *The Cross-Language Evaluation Forum 2006 Working Notes*, Alicante, Spain (2006)
3. Koza, J.R.: *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Complex Adaptive Systems, MIT Press, Cambridge 840 (1992)
4. Porter, M.F: An algorithm for suffix stripping. *Program* 14(3), 130–137 (1980)
5. Salton, G.: *The SMART Information Retrieval System*. Prentice Hall, Englewood Clis (1971)
6. Automatic text processing: the transformation, analysis, and retrieval of information by computer, p. 530. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA (1999)

GEOUJA System. The First Participation of the University of Jaén at GEOCLEF 2006

Manuel García-Vega¹, Miguel A. García-Cumbreras¹, L.A. Ureña-López¹,
and José M. Perea-Ortega¹

Dpto. Computer Science. University of Jaén.Spain
{mgarcia,magc,laurena,jmperea}@ujaen.es

Abstract. This paper describes the first participation of the SINAI group of the University of Jaén in GeoCLEF 2006. We have developed a system made up of three main modules: the Translation Subsystem, that works with queries into Spanish and German against English collection; the Query Expansion subsystem, that integrates a Named Entity Recognizer, a thesaurus expansion module and a geographical information-gazetteer module; and the Information Retrieval subsystem. We have participated in the monolingual and the bilingual tasks. The results obtained shown that the use of geographical and thesaurus information for query expansion does not improve the retrieval in our experiments.

1 Introduction

The objective of GeoCLEF is to evaluate Geographical Information Retrieval (GIR) systems in tasks that involves both spatial and multilingual aspects. Given a multilingual statement describing a spatial user need (topic), the challenge is to find relevant documents from target collections into English, using topics into English, Spanish, German or Portuguese [1,2].

The main objective of our first participation in GeoCLEF have been the study of the problem of this task, and to develop a system to solve some of them. For this reason, our system consist of three subsystems: Translation, Query Expansion and Information Retrieval. The Query Expansion subsystem is formed as well by three modules: Named Entity Recognition, Geographical Information-Gazetteer and Thesaurus Expansion.

Next section describes the whole system and each module of the system. Then, in the section 3 experiments and results are described. Finally, the conclusions about our first participation in GeoClef 2006 are expounded.

2 System Description

We propose a Geographical Information Retrieval System that is made up of three subsystems (See Figure 1):

- **Translation Subsystem:** is the query translation module. This subsystem translates the queries to the other languages. For the translation an own module has been used, called SINTRAM (SINai TRANslation Module)¹, that works with several online Machine Translators, and implements several heuristics. For these experiments we have used an heuristic that joins the translation of a default translator (the one that we indicate depends of the pair of languages), with the words that have another translation (using the other translators).
- **Query Expansion Subsystem:** the goal of this subsystem is to expand the queries with geographical data and words from the thesaurus (see below). It is made up of three modules: a Named Entity Recognition Module, that uses a Geographical Information-Gazetteer Module, and a Thesaurus Expansion Module. These modules are described in detail next:
 - **Named Entity Recognition (NER) Module:** the main goal of NER Subsystem is to detect and recognize the *location* entities in the queries, in order to expand the topics with geographical data (using the Geographical Information Module). We have used the NER module of GATE² and its own Gazetteer. The location terms includes everything that is town, city, capital, country and even continent. The NER Module generates some labelled topics adding the found locations.
 - **Geographical Information Module:** This module stores the geographical data. This information has been obtained from the Geonames database³. We have used it only for English because previously all the queries have been translated. The goal of this module is to expand the locations of the topics recognized by the NER module, using geographical information. The have made automatic query expansion[3]. When a location is recognized by the NER Module the system look for in the Geographical Information Module. In addition, it is necessary to consider the spatial relations found in the query (“near to”, “within X miles of”, “north of”, “south of”, etc.). Depending on the spatial relations, the search in the Geographical Information Module is more or less restrictive.
 - **Thesaurus Expansion Module:** This is the query expansion module, using an own thesaurus. A collection of thesauri was generated from the GeoCLEF English training corpus. This module was looking for words with a very high rate of document co-location, using the standard *TF.IDF* for words comparing test. These words were treated like synonyms and added to the topics. For that, we generated an inverse file with the GeoCLEF 2005 corpus. The file has a row for each different word of the corpus with the word frequencies for each corpus file. We found that a cosine similarity great than 0.9 between words was the rate

¹ <http://sinai.ujaen.es>

² <http://gate.ac.uk/>

³ <http://www.geonames.org/>. Geonames geographical database contains over eight million geographical names and consists of 6.3 million unique features whereof 2.2 million populated places and 1.8 million alternate names.

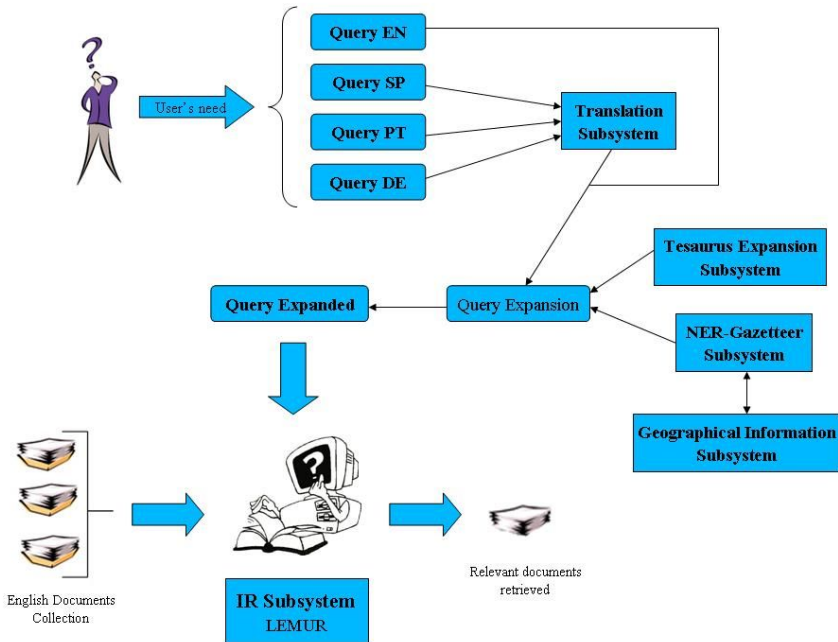


Fig. 1. GEOUJA System architecture

that obtain best precision/recall results (in average 2 words added). The same procedure was applied to the 2006 corpus.

- **Information Retrieval Subsystem:** We have used LEMUR IR system⁴.

3 Results

In our **baseline** experiment we have used the original English topics set. All topics are preprocessed (stopper and stemmer) without query expansion (geographical or thesaurus). We have used Okapi as weighting function and pseudo-relevant feedback (PRF) in all experiments.

We have participated in monolingual and bilingual tasks. In monolingual task with five experiments (the official results are shown in Table 1):

- *sinaiEnEnExp1*. Baseline experiment, considering all tags (title, description and narrative). Without expansion topics.
- *sinaiEnEnExp2*. The same experiment as baseline but considering only title and description tags.
- *sinaiEnEnExp3*. We have considered only title and description tags. The query has been expanded using the title with geographical information.

⁴ <http://www.lemurproject.org/>

- *sinaiEnEnExp4*. We have considered only title and description tags. The query has been expanded using the title and the description with thesaurus information.
- *sinaiEnEnExp5*. We have considered only title and description tags. The query has been expanded using the title and the description with geographical and thesaurus information.

For the bilingual task we have made five experiments: two experiments for German-English task and three experiments for Spanish-English task (the official results are shown in Table 2):

- *sinaiDeEnExp1*. All the query tags have been translated from German to English and preprocess (stopper and stemmer), and the query has not been expanded.
- *sinaiDeEnExp2*. The same experiment as previous but only considering the title and description tags.
- *sinaiEsEnExp1*. All the query tags have been translated from Spanish to English and preprocess (stopper and stemmer), and the query has not been expanded.
- *sinaiEsEnExp2*. The same experiment as previous but only considering the title and description tags.
- *sinaiEsEnExp3*. We have considered only title and description tags. The query has been expanded using the title with geographical information.

Table 1. Official results in monolingual task

Experiment	Mean Average Precision	R-Precision
sinaiEnEnExp1	0.3223	0.2934
sinaiEnEnExp2	0.2504	0.2194
sinaiEnEnExp3	0.2295	0.2027
sinaiEnEnExp4	0.2610	0.2260
sinaiEnEnExp5	0.2407	0.2094

Table 2. Official results in bilingual task

Experiment	Mean Average Precision	R-Precision
sinaiDeEnExp1	0.1868	0.1649
sinaiDeEnExp2	0.2163	0.1955
sinaiEsEnExp1	0.2707	0.2427
sinaiEsEnExp2	0.2256	0.2063
sinaiEsEnExp3	0.2208	0.2041

4 Conclusions

The obtained results shown that the way we have implemented query expansion (using geographical and thesaurus information) did not improve the retrieval.

Several reasons exist to explain the worse results obtained with the expansion of topics, and these are our conclusions:

- The NER module used sometimes does not work well, because in some topics only a few entities are recognized and not all. For the future we will test other NERs.
- In the topics, sometimes appear compound locations like New England, Middle East, Eastern Bloc, etc., that not appear in the Geographical Information-Gazetteer Module.
- Depending on spatial relation in the topics, we could improve the expansion, testing in which cases the system works better with more locations or less. Therefore, we will try to improve the Geographical Information-Gazetteer Module and the Thesaurus Expansion Module to obtain better query expansions.

As future work we want to know why the expansion module did not work as well as we expected. It is known that sometimes the Geonames module introduce noise in the queries, but our thesauri should improve the baseline method. We also want to include in the system another module that expand the queries using Google.

Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

References

1. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks. In: Working Notes for the CLEF 2006 Workshop (2006)
2. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 Medical Retrieval and Annotation Tasks. In: Working Notes for the CLEF 2006 Workshop (2006)
3. Buscaldi, D., Rosso, P., Sanchis-Arnal, E.: A WordNet-based Query Expansion method for Geographical Information Retrieval. In: Working Notes for the CLEF 2005 Workshop (2005)

R2D2 at GeoCLEF 2006: A Combined Approach

Manuel García-Vega¹, Miguel A. García-Cumbreras¹
L. Alfonso Ureña-López¹, José M. Perea-Ortega¹, F. Javier Ariza-López¹
Oscar Ferrández², Antonio Toral², Zornitsa Kozareva²
Elisa Noguera², Andrés Montoyo², Rafael Muñoz²
Davide Buscaldi³, and Paolo Rosso³

¹ University of Jaén

{mgarcia,magc,laurena,jmperea,fjariza}@ujaen.es

² University of Alicante

{ofe,atoral,zkozareva,elisa,montoyo,rafael}@dlsi.ua.es

³ Polytechnical University of Valencia

{dbuscaldi,proso}@dsic.upv.es

Abstract. This paper describes the participation of a combined approach in GeoCLEF-2006. We have participated in Monolingual English Task and we present joint work of the three groups or teams belonging to the project R2D2¹ with a new system, combining the three individual systems of these teams. We consider that research in the area of GIR is still in its very early stages, therefore, although a voting system could improve the individual results of each system, we have to further investigate different ways to achieve a better combination of these systems.

1 Introduction

GeoCLEF is a track of the Cross-Language Evaluation Forum (CLEF) whose aim is to provide the necessary framework in which to evaluate Geographic Information Retrieval (GIR) Systems for search tasks involving both spatial and multilingual aspects. The relevant documents are retrieved by using geographic tags like geographic places, geographic events and so on. Nowadays, the fast development of Geographic Information Systems (GIS) involves the need of GIR system that helps these systems to obtain documents with relevant geographic information.

In this paper we present joint work of the three groups belonging to the project R2D2: UJA², UA³ and UPV⁴.

¹ In Spanish “Recuperación de Respuestas en Documentos Digitalizados” —Answering Retrieval in Digitized Documents—. Financed project by Spanish Government with grant TIC2003-07158-C04.

² University of Jaén with RIM subproject and reference TIC2003-07158-C04-04.

³ University of Alicante with BRUJULA subproject and reference TIC2003-07158-C04-04-01.

⁴ Polytechnical University of Valencia with SIRICM subproject and reference TIC2003-07158-C04-03.

Our approach consists of a combined system based on the three systems' scores of the aforementioned teams. These scores are treating individually, generating a new one by means of voting strategy. Since this is the first year of the voting system, we have used a simple method.

The rest of the paper is organized as follows: Sections 2, 3 and 4 describes the individual systems, firstly the UJA system, secondly the UA system and finally the UPV system. Section 5 describes the voting system and section 6 shows the results. Finally, section 7 illustrates some conclusions and future work proposals.

2 Description of the System of the University of Jaén

The SINAI⁵ team at University of Jaén propose a Geographical Information Retrieval System that is composed of five subsystems:

- **Translation Subsystem:** is the query translation module. This subsystem translates the queries to the other languages and it is used for the following bilingual tasks: Spanish-English, Portuguese-English and German-English. For the translation, a dedicated module has been used, called SINTRAM (SINai TRAnslation Module), that works with several online machine translation services, and implements several heuristics. For these experiments we have used an heuristic that joins the translation of a default translator (the one that we indicate depends of the pair of languages), with the words that have been alternatively translated by another translation service (using the other translators).
- **Named Entity Recognition-Gazetteer Subsystem:** is the query geo-expansion module. The main goal of the NER-Gazetteer Subsystem is to detect and recognize the entities in the queries, in order to expand the topics with geographical information. We are only interested in geographical information, so we have used only the locations detected by the NER module. The *location* term includes everything that is *town*, *city*, *capital*, *country* and even *continent*. The information about locations is loaded previously in the Geographical Information Subsystem, that is related directly to the NER-Gazetteer Subsystem. The NER-Gazetteer Subsystem generates labels the original topics by adding the locations found.
- **Geographical Information Subsystem:** is the module that stores the geographical data. This information has been obtained from the Geonames⁶ gazetteer. The objective of this module is to expand the locations of the topics, using geographical information. The expansion that we do is the *automatic query expansion*[2]. The Geonames database contains over six million entries for geographical names, whereof 2.2 million are cities and villages. It integrates geographical data such as names, altitude, population and others from various sources. When a location is recognized by the NER subsystem we look for in the Geographical Information Subsystem. In addition, it is

⁵ <http://sinai.ujaen.es/>

⁶ <http://www.geonames.org/>

necessary to consider the spatial relations found in the query (“near to”, “within X miles of”, “north of”, “south of”, etc.). Depending on the spatial relations, the search in the Geographical Information subsystem is more or less restrictive.

- **Thesaurus Expansion Subsystem:** is the query expansion module using a thesaurus developed at the University of Jaén. A collection of thesauri was generated from the GeoCLEF training corpus. This subsystem was looking for words with a very often cooccurring in documents. These words were treated like synonyms and added to the topics. An inverse file with the entire collection was generated for comparing words. Training with GeoCLEF-2005 files, a cosine similarity of 0.9 turned out to be the best parameter setting based on the recall and precision of the results.
- **IR Subsystem:** is the Information Retrieval module. The English collection dataset has been indexed using the LEMUR IR system⁷. It is a toolkit⁸ that supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or subcollections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

3 Description of the System of the University of Alicante

The aim of the University of Alicante approach is to evaluate the impact of the application of geographic knowledge extracted from a structured resource over a classic Information Retrieval (IR) system. In the Figure 2, an overview of the system is depicted.

The GIR system developed by the University of Alicante for the second edition of GeoCLEF is made up of two main modules. Here we show a brief description of them, for further details see [5].

IR. A Passage Retrieval (PR) module called IR-n [3] has been used for several years in the GeoCLEF campaign. It allows using different similarity measures. The similarity measure used has been dfr as it has been the one that obtained the best results when trained over CLEF corpora.

Geographic knowledge. Also, a geographic database resource called Geonames⁹ was used, which is freely available and may be used through web services or downloaded as a database dump. It is a structured resource built from different sources such as NGA, GNIS and Wikipedia. For each entry it contains several fields which provide name, additional translations, latitude, longitude, class according to a self-developed taxonomy, country code, population and so on. This module is used to enrich the initial information provided by topics with related geographic items.

⁷ <http://www.lemurproject.org/>

⁸ The toolkit is being developed as part of the Lemur Project, a collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

⁹ <http://www.geonames.org>

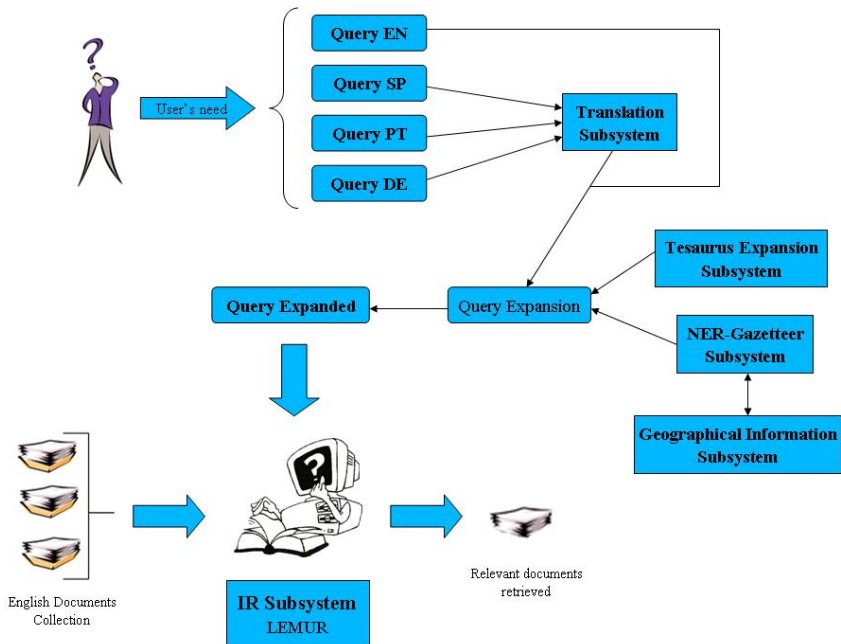


Fig. 1. UJA system architecture

The system works in the following way:

1. Topic processing: the topic is processed in order to obtain the relevant geographic items and the geographical relations among them. Besides, all the nouns in the topic which are not of a geographic nature are marked as required words, belonging to topics widely used words (e.g. document) nor stop words.
2. Geographic query: once the geographic information from the topic is obtained, a query to the Geonames database is methodically built. This has the aim of obtaining related geographic items from this resource. In order to build this query information, longitude, latitude or country names are considered.
3. Topic enrichment: a new topic is composed of the information contained in the provided topic and the geographic items obtained in the previous step.
4. Information Retrieval: finally, the relevant documents from the collection according to the topic and its related geographic items are obtained. For a document to be considered in the retrieval, it should contain at least one of the required words. The objective of this is to lessen the noise that could be introduced by adding big lists of geographic items to the topic.

Even though we have incorporated geographic knowledge in our system and the run enriched with this knowledge was used for the combined system, we can conclude that the research in GIR is at the very first steps. This claim is

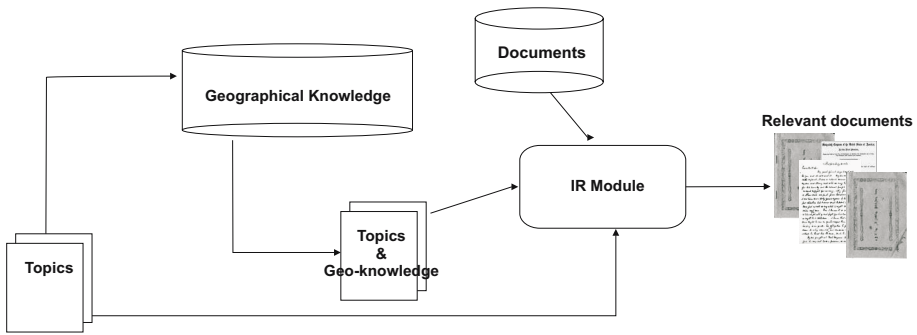


Fig. 2. UA system architecture

supported by the fact that the systems that obtained the best results in the first edition of GeoCLEF were the ones using classic IR without any geographic knowledge. Therefore, we plan to continue this line of research, whose principal aim is to make out how to incorporate information of this nature so that the systems can benefit from it.

4 Description of the System of the Polytechnical University of Valencia

The GeoCLEF system of the UPV is based on a WordNet-based expansion of the geographical terms in the documents, that exploits the synonymy and holonymy relationships. This can be seen as an “inverse” approach with respect to the UPV’s 2005 system [1] which exploited the meronymy and synonymy in order to perform a query expansion. Query expansion was abandoned due to the poor results obtained in the previous edition of the GeoCLEF.

WordNet [4] is a general-domain ontology, but includes some amount of geographical information that can be used for the Geographical Information Retrieval task. However, it is quite difficult to calculate the number of geographical entities stored in WordNet, due to the lack of an explicit annotation of the synsets. We retrieved some figures by means of the *has_instance* relationship, resulting in 654 cities, 280 towns, 184 capitals and national capitals, 196 rivers, 44 lakes, 68 mountains. As a comparison, a specialized resource like the Getty Thesaurus of Geographic Names (TGN)¹⁰ contains 3094 entities of type “city”.

The indexing process is performed by the Lucene¹¹ search engine, generating two index for each text: a *geo* index, containing all the geographical terms included in the text together with those obtained by means of WordNet, and a *text* index, containing the stems of text words that are not related to geographical entities only. Thanks to the separation of the indices, a document containing “John Houston” will not be retrieved if the query contains “Houston”, the city

¹⁰ http://www.getty.edu/research/conducting_research/vocabularies/tgn/

¹¹ <http://lucene.apache.org>

in Texas. The adopted weighting scheme is the usual *tf-idf*. The geographical names were detected by means of the Maximum Entropy-based tool available from the *openNLP* project¹².

Since the tool does not perform a classification of the named entities, the following heuristics is used in order to identify the geographical ones: when a Named Entity is detected, we look in WordNet if one of the word senses has the *location* synset among its hypernyms. If this is true, then the entity is considered a geographical one.

For every geographical location *l*, the synonyms of *l* and all its holonyms (even the inherited ones) are added to the *geo* index. For instance, if *Paris* is found in the text, its synonyms *City of Light*, *French capital*, *capital of France* are added to the *geo* index, together with the holonyms: {*France*, *French republic*}, {*Europe*}, {*Eurasia*}, {*Northern Hemisphere*} and {*Eastern Hemisphere*}. The obtained holonyms tree is:

```
Paris, City of Light, French capital
=>France, French republic
=>Europe
=>Eurasia
=>Northern Hemisphere
=>Eastern Hemisphere
```

The advantage of this method is that knowledge about the enclosing, broader, geographical entities is stored together with the index term. Therefore, any search addressing, for instance, *France*, will match with documents where the names *Paris*, *Lyon*, *Marseille*, etc. appear, even if *France* is not explicitly mentioned in the documents.

5 The Combined System

These three systems have returned their own final scores. Our approach has been a combined system based on the individual scores generating a new system with voted final results.

Since this is the first year of the voting system, we have used the simplest voting method:

1. The systems have their own scoring method and we need a normalized version of them for a correct fusion. After this procedure, all scores will have values between 0 and 1.
2. If a document is very well considered in the three systems, we want it to have a good value in the combined one. We have used the addition of the normalized values as the final score. If a document appears in more than one system, it has a score result of the sum of each score.
3. A list of new scores was generated from the final results of all systems. Finally, we obtained our final score by sorting this list and cutting in the 1,000th position.

¹² <http://opennlp.sourceforge.net>

6 Experimental Results

Our combined approach has only participated in monolingual task and the official results are shown in Table 1 and Table 2. The results shown that the combined approach must be improved because the average precision obtained is poor. This could be done analyzing the advantages of each system in each case or topic and to obtain what system works better for each spatial relation or type of region (location, country, etc.). The future system of voting would have to consider the previous analysis and to weigh with greater or smaller score the results of each system depending on the type of question or spatial relation.

Table 1. Average precision in monolingual task

Interpolated Recall (%)	Precision Averages (%)
0%	43,89%
10%	36,13%
20%	34,85%
30%	33,53%
40%	33,09%
50%	32,25%
60%	24,06%
70%	14,30%
80%	11,60%
90%	7,30%
100%	6,08%
Average precision	24,03%

Table 2. R-precision in monolingual task

Docs Cutoff Levels	Precision at DCL (%)
5 docs	21,60%
10 docs	17,60%
15 docs	16,80%
20 docs	16,60%
30 docs	13,73%
100 docs	7,40%
200 docs	4,42%
500 docs	2,09%
1.000 docs	1,25%
R-Precision	23,19%

7 Conclusions and Future Work

We have presented a combined approach using other three GeoCLEF-2006 systems. Although the results have not been very high we consider them quite

promising. The combined approach takes advantages of them, although it acquires their defects too. The linearity of the voting system biases documents common in all three individual list of documents whether relevant or not.

As future work we intend to improve the combined approach using three individual systems considering other voting strategies. Using some techniques from artificial intelligence for detecting the quality of the individual systems could improve the final results. A neural network can tune the precision of the individual system for improve the behavior of the combined one.

Acknowledgments

This work has been supported by Spanish Government with grant TIC2003-07158-C04-04 (University of Jan), TIC2003-07158-C04-04-01 (University of Alicante) and TIC2003-07158-C04-03 (Polytechnical University of Valencia).

References

1. Buscaldi, D., Rosso, P., Sanchis, E.: Using the wordnet ontology in the geoclef geographical information retrieval task. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 939–946. Springer, Heidelberg (2006)
2. Buscaldi, D., Rosso, P., Sanchis-Arnal, E.: A wordnet-based query expansion method for geographical information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
3. Llopis, F.: IR-n un Sistema de Recuperacin de Informacin Basado en Pasajes. *Procesamiento del Lenguaje Natural* 30, 127–128 (2003)
4. Miller, G.A.: Wordnet: A lexical database for english. *Communications of the ACM* 38, 39–41 (1995)
5. Toral, A., Ferrández, O., Noguera, E., Kozareva, Z., Montoyo, A., Muñoz, R.: Geographic IR Helped by Structured Geospatial Knowledge Resources. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) *Working Notes at CLEF 2006* (2006)

MSRA Columbus at GeoCLEF 2006

Zhisheng Li¹, Chong Wang², Xing Xie², Xufa Wang¹, and Wei-Ying Ma²

¹Department of Computer Science and Technology, University of Sci. & Tech. of China,
Hefei, Anhui, 230026, P.R. China
zsli@mail.ustc.edu.cn, xfwang@ustc.edu.cn

²Microsoft Research Asia,
4F, Sigma Center, No.49, Zhichun Road, Beijing, 100080, P.R. China
{chwang, xingx, wyma}@microsoft.com

Abstract. This paper describes the participation of Columbus Project of Microsoft Research Asia (MSRA) in GeoCLEF 2006. We participated in the Monolingual GeoCLEF evaluation (EN-EN) and submitted five runs based on different methods. In this paper, we describe our geographic information retrieval system, discuss the results and draw following conclusions: 1) if we just extract locations from the topics automatically without any expansions as geo-terms, the retrieval performance is barely satisfactory; 2) automatic query expansion weakens the performance; 3) if the queries are expanded manually, the performance is significantly improved.

Keywords: Location extraction and disambiguation, geographic focus detection, implicit location, geographic information retrieval.

1 Introduction

In common web search and mining, location information is usually discarded. However, people need to utilize locations for many purposes, such as dining traveling and shopping. The goal of the Columbus Project is to utilize location information in web search and mining to better facilitate users to acquire the knowledge related to locations and suggest a location-aware way to organize the information.

This is the first time for Columbus project to participate in GeoCLEF 2006 [1]. GeoCLEF aims at providing a necessary platform for evaluating geographic information retrieval systems. We participated in the Monolingual GeoCLEF evaluation (EN-EN) and submitted five runs based on different methods.

2 Geographic Information Retrieval System: An Overview

Our geographic information retrieval system is composed of five modules, including geographic knowledge base, location recognition and disambiguation module, geographic focus detection module, geo-indexing module and geo-ranking module. More details about these modules can be found in [2]. We index the collections provided by the organizers by both text-index and geo-index. Afterwards, we translate the topics from GeoCLEF 2006 into queries in our input format based on different

methods. Finally, we rank the results by considering both text relevance and geo relevance.

3 Monolingual GeoCLEF Experiments (English - English)

In Table 1 we show all the five runs submitted to GeoCLEF. When the topic field is “Title”, we just use the EN-title element of the topic to generate the query of the run. When the topic field is “Title + Description”, this means that both the EN-title and EN-desc are used in the run. When the topic field is “Title + Description + Narrative”, this means that EN-title, EN-desc and EN-narr are all used. Priorities are assigned by us, where priority 1 is the highest and 5 the lowest.

Table 1. Run information

Run-ID	Topic Fields	Priority
MSRAWhitelist	Title	1
MSRAManual	Title + Description	2
MSRAExpansion	Title + Description	3
MSRALocal	Title	4
MSRAText	Title + Description + Narrative	5

In MSRAWhitelist, we use the EN-title elements of the topics to generate the queries. For some special queries, e.g. “Credits to the former Eastern Bloc”, “Arms sales in former Yugoslavia”, “Eastern Bloc” and “former Yugoslavia” are not in our gazetteer, so we utilize the geo knowledge base to get the corresponding sovereigns of these geo-entities. Then we can make a white list manually for the geo-terms of these queries.

In MSRAManual, we generate the queries with EN-title and EN-desc elements of the topics. However for some queries, e.g. “Wine regions around rivers in Europe”, “Tourism in Northeast Brazil”, if we just input the “wine regions” or “Tourism”, very few documents return. So we manually modify the textual terms of such queries, e.g. “Tourism” -> “tourism tourist tour traveling travel”.

In MSRAExpansion, we generate the queries with EN-title and EN-desc elements of the topics. Different from MSRAManual, the queries are automatically expanded based on the pseudo-feedback technique. First we use the original queries to search the corpus. Then we extract the locations from the returned documents and calculate the times each location appears in the documents. Finally we get the top 10 most frequent location names and combine them with the original geo-terms as new queries. For example, “former Yugoslavia” is expanded to be “Croatia, United States, Russia, Sarajevo, Balkan, Edinburgh, Tuzla, Belgrade, Germany, Zagreb”. Then we can see “Croatia, Sarajevo, Belgrade, Zagreb, Tuzla” are really parts of “former Yugoslavia”, however, unrelated locations, such as “United States”, also appears. The reason is that they appear together with “former Yugoslavia” frequently in the corpus and it’s difficult to find a criterion to remove them.

In MSRALocal, we use the EN-title elements of the topics to generate the queries. And we do not use geo knowledge base or query expansion method to expand the

query locations. We just utilize our location extraction module to extract the locations automatically from the queries.

In MSRAText, we generate the queries with EN-title, EN-desc and EN- narr elements of the topics. We just utilize our pure text search engine to process the queries.

4 Results and Discussion

Fig.1 shows the interpolated recall vs average precision for the five runs. MSRAManual is the best run among the five ones as expected, because the queries are modified manually. The worst one is MSRAExpansion. The reason is that many unrelated locations are added to new queries after pseudo feedback for some topics.

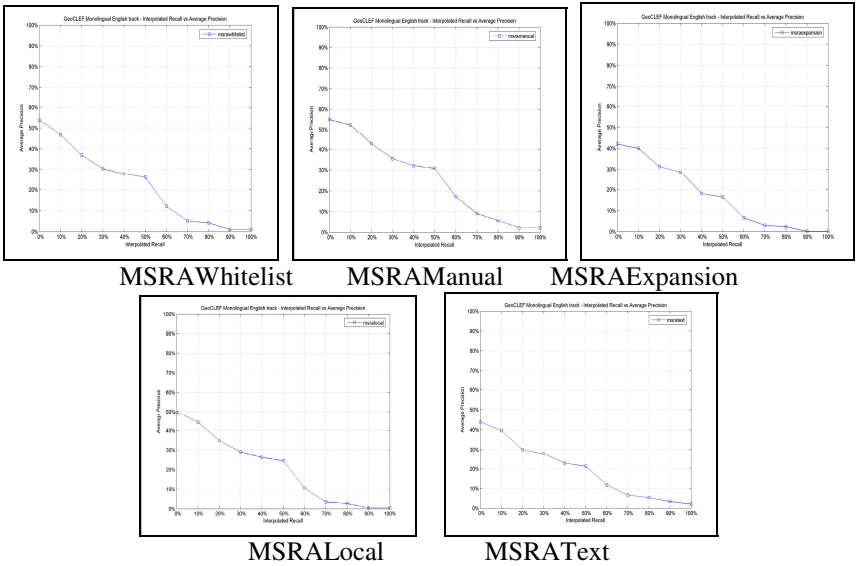


Fig. 1. The interpolated recall vs average precision for all five runs

Table 2 shows the MAP and R-Precision for all the runs. MSRAManual outperforms the other runs in both MAP and R-Precision. The performances for MSRALocal and MSRAText are very similar to each other. It indicates that only extracting locations from the topics as geo-terms without any expansion won’t improve MAP much. However, MSRAWhitelist outperforms MSRALocal by about 1.7% in MAP. The reason is that we only made white list for several topics and it won’t affect the MAP much.

MSRAManual improves the performance significantly compared with MSRALocal. For example, for topic “Tourism in Northeast Brazil”, the MAP of MSRALocal is 0.0%, but in MSRAManual it’s 26.96%. This is because, as discussed in Section 3, very few documents contain the keyword “Tourism”. We manually

modify “Tourism” to “tourism tourist tour traveling travel” and improve the performance for this query significantly. For topic “Archeology in the Middle East”, the MAP increases from 6.45% to 27.31%, and the reason is similar.

The MAP of MSRAExpansion drops a little compared with MSRALocal, since the expansion introduces unrelated locations. Perhaps, if we have a large corpus, the performance of the expansion approach would improve, since the top-ranked documents might contain more relative locations.

Table 2. Map & R-Precision for five runs

RUN-ID	MAP	R-Precision
MSRAWitelist	20.00%	23.52%
MSRAManual	23.95%	25.45%
MSRAExpansion	15.21%	18.53%
MSRALocal	18.37%	22.45%
MSRAText	18.35%	21.23%

5 Conclusion

Our experimental results indicate that simply extracting locations without query expansion does not lead to satisfactory retrieval performance. On the other hand, fully automatic query expansion also weakens the performance. The reason is that the topics are usually too short to understand and the corpus is not large enough for query expansion. One possible solution is to perform the query expansion using larger corpus such as the Web. In addition, if the queries are expanded manually, the performance can be significantly improved.

Our future work include: 1) efficient geographic indexing and ranking approaches; 2) automatic query parsing; 3) geographic disambiguation algorithms and 4) result visualization.

References

1. GeoCLEF 2006, <http://ir.shef.ac.uk/geoclef/>
2. Li, Z.S., Wang, C., Xie, X., Wang, X.F., Ma, W.Y.: Indexing implicit locations for geographic information retrieval. In: GIR’06, Seattle, USA,

Forostar: A System for GIR

Simon Overell¹, João Magalhães¹, and Stefan Rüger^{1,2}

¹ Multimedia and Information Systems

Department of Computing, Imperial College London SW7 2AZ, UK

² Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK
{simon.overell, j.magalhaes}@imperial.ac.uk, s.rueger@open.ac.uk

Abstract. We detail our methods for generating and applying co-occurrence models for the purpose of placename disambiguation. We explain in detail our use of co-occurrence models for placename disambiguation using a model generated from Wikipedia. The presented system is split into two stages: a batch text & geographic indexer and a real time query engine. Four alternative query constructions and six methods of generating a geographic index are compared. The paper concludes with a full description of future work and ways in which the system could be optimised.

1 Introduction

In this paper we detail *Forostar*, our GIR system designed to enter GeoCLEF 2006. We begin with a full outline of the system followed by our experimental runs. We aim to test the accuracy of our co-occurrence model and how the use of large scale co-occurrence models can aid the disambiguation of geographic entities. We conclude with an analysis of our results and future work.

We use a *rule-based* approach to annotate how placenames occur in Wikipedia (taking advantage of structure and meta-data). This annotated corpus is then applied as a co-occurrence model using a *data-driven* method to annotate the GeoCLEF data.

1.1 Discussion on Ambiguity

GeoCLEF is an appropriate framework for developing different methods of placename disambiguation, however, evaluation is difficult as no ground truth exists for the corpus. The method we use for indexing locations is a *unique geographic index*, every occurrence of a placename is represented as a single polygon on the earth's surface in a spatial index. This allows for efficiently comparing the locations referred to in a query to documents in the corpus.

Despite there being minimal ambiguity in the GeoCLEF queries themselves, due to the use of a unique geographic index we would still expect to see an improvement in MAP as the accuracy of the index increases. This is because locations not necessarily occurring in the query but appearing in the footprint described by the query being incorrectly classified causing false negatives.

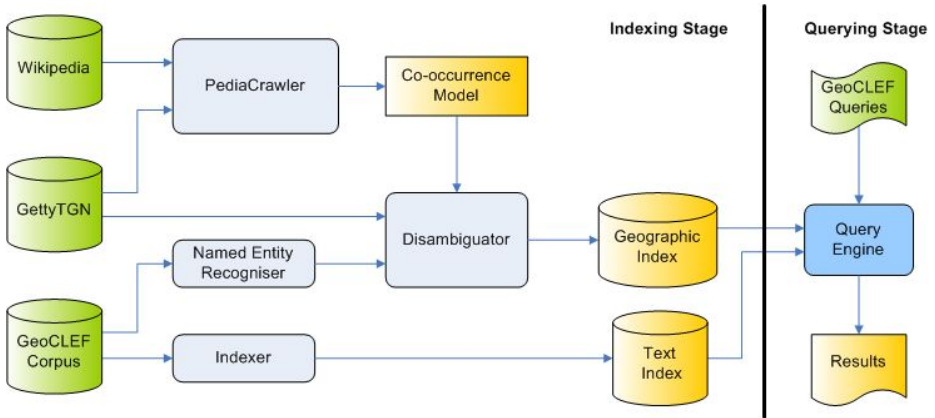


Fig. 1. System Design

2 The System

Forostar is split into two parts: the *indexing stage* and the *querying stage* (Figure 1). The indexing stage requires the corpus and some external resources to generate the geographic and text indexes (a *slow* task). The querying stage requires the generated indexes and the queries; it runs in *real time*.

The Indexing stage consists of four separate applications: *PediaCrawler* is first used to crawl the links in Wikipedia building a co-occurrence model of how placenames occur [9]; *Disambiguator* then applies the co-occurrence model to disambiguate the named entities extracted from the GeoCLEF corpus by the *Named Entity Recogniser* [2]. The disambiguated named entities form the geographic index. *Indexer* is used to build the text index.

The Querying stage consists of our *Query Engine*; this queries the text and geographic indexes separately, combining the results.

2.1 PediaCrawler

PediaCrawler is the application designed to build our co-occurrence model. Wikipedia articles and stubs are crawled¹; articles referring to placenames are mapped to locations in the Getty Thesaurus of Geographic Names (TGN). Our co-occurrence model takes the form of two database tables: the Mapping table, a mapping of Wikipedia articles to TGN unique identifiers; and the Occurrences table, links to articles believed to be places and the order in which they occur, no other information is used in the model.

PediaCrawler uses rule-based methods of disambiguation. It is made up of two parts, the *disambiguation framework* and a *method of disambiguation*. By using Wikipedia to build our co-occurrence model we hope to solve two problems: the

¹ Our copy of Wikipedia was taken 3rd Dec 2005.

problem of *synonyms* (multiple placenames referring to a single location) is resolved by recording how multiple anchor texts point to the same page; and the problem of *polynyms* (a single placename referring to multiple locations) can be solved with our disambiguation system.

The disambiguation framework. The disambiguation framework is a simple framework to abstract the method of disambiguation from the document crawler. The framework is outlined as follows: *PediaCrawler* loads the Wikipedia articles to be crawled from a database, the links from each article are crawled in turn. For each link, if the article linked to has already been classified update the Occurrences table, otherwise, classify the article using the *Method of Disambiguation* specified and update both the Occurrences and Mapping tables.

The disambiguation method uses the following information to disambiguate a link: a set of candidate locations; a list of related placenames extracted from the metadata in the article pointed to by the link; and the text and title of the article pointed to by the link. Where the candidate locations are the set of matching locations found in the TGN by considering the link and title as placenames.

Our method of disambiguation. Based on the results observed by running a series of simple disambiguation methods on test data, we designed a disambiguation pipeline that could exploit the meta-data contained in Wikipedia and balance precision and recall (maximising the F_1 measure) [9].

Each disambiguation pipeline step is called in turn. A list of candidate locations is maintained for each article, an article is denoted as unambiguous when this list contains one or zero elements. Each method of disambiguation can act on the candidate locations list in the following ways: remove a candidate location; add a candidate location; remove all candidate locations (disambiguate as not-a-location); or remove all but one candidate locations (disambiguate as a location).

1. Disambiguate with templates – The template data in Wikipedia is highly formatted data contained in name-value pairs. The format of the templates is as follows: $\{\{template\ name\ |\ name_1 = value_1\ |\ \dots\ |\ name_i = value_i\}\}$. The template name is initially used for disambiguation, for example “Country” will indicate this page refers to a location of feature type nation or country. Templates are also used to identify non-places, for example if the template type is “Biographic” or “Taxonomic.” The name-value pairs within a template are also used for disambiguation, e.g. in the Coord template a latitude and longitude are provided which can be matched to the gazetteer.

2. Disambiguate with categories – The category information from Wikipedia contains softer information than the template information [7]; the purpose of categorising articles is to denote associations between them (rather than templates which are intended to display information in a uniform manner). Category tags can identify the country or continent of an article, or indicate an article is not referring to a location.

3. Disambiguate with co-referents – Often in articles describing a location, a parent location will be mentioned to provide reference (e.g. when describing a town, mention the county or country). The first paragraph of the document

is searched for names of containing locations. This method of disambiguation has been shown to have a high location-precision (articles correctly identified as locations) and grounding (articles correctly matched to unique identifiers), 87% and 95% respectively [9].

4. Disambiguate with text heuristics – Our heuristic method is based on the hypothesis: *When describing an important place, only places of equal or greater importance are used to provide appropriate reference.*

This hypothesis led to the implementation of the following disambiguation procedure: all the placenames are extracted from the first paragraph of the document; for each possible location of the ambiguous placename, sum the distance between the possible location and the extracted locations with a greater importance; classify as the location with the minimal sum.

2.2 Named Entity Recogniser

News articles have a large number of references to named entities that quickly place the article in context. The detection of references to all named entities is the problem addressed in this part of the system. The named entity recogniser receives as input the GeoCLEF news articles and outputs the named entities of each article.

Named entity recognition systems rely on lexicons and textual patterns either manually crafted or learnt from a training set of documents. We used the E-Spotter named entity recognition system proposed by Zhu et al. [14]. Currently, ESpotter recognises people, organisations, locations, research areas, email addresses, telephone numbers, postal codes, and other proper names. First it infers the domain of the document (e.g. computer science, sports, politics) to adapt the lexicon and patterns for a more specialised named entity recognition which will result in a high precision and recall.

ESpotter uses a database to store the lexicon and textual pattern information; it can easily be customised to recognise any type of entity. The database we used is the one supplied by Zhu et al., we did not create a GeoCLEF specific database.

2.3 Indexer

The news article corpus was indexed with Apache Lucene 2.0 [1], later also used to search the article corpus. The information retrieval model used was the vector space model without term frequencies (binary term weight). This decision was due to the small size of each document that could cause a large bias for some terms. Terms are extracted from the news corpus in the following way: split words at punctuation characters (unless there is a number in the term); recognise email addresses and internet host names as one term; remove stop words; index a document by its extracted terms (lowercase) (see [1] for details).

2.4 Disambiguator

The *Disambiguator* builds a geographic index allowing results from a text search to be re-ordered based on location tags. The named entities tagged as placenames

output by the *Named Entity Recogniser* are classified as locations based on their context in the co-occurrence model. This applies Yarowsky's "One sense per collocation" property [12].

The geographic index is stored in a Postgres database and indexed with an R-Tree to allow efficient processing of spatial queries [3]. In previous experiments we have shown the co-occurrence model to be accurate up to 80% [9], in this experiment we assume the geographic index to have an accuracy equal to or less than this.

Disambiguation methods: We compared three base-line methods of disambiguation to three more sophisticated methods:

No geographic index (NoGeo). In this method of disambiguation no geographic index is used. As with traditional IR methods, only the text parts of the query are executed. The motivation of this method is to measure to what extent the geographic index affects the results.

No disambiguation (NoDis). In this method no disambiguation is performed. Each placename occurrence is indexed multiple times, once for each matching location in the co-occurrence model. For example if "Africa" appears in a document it will be added to the geographic index twice: (Africa – Continent) and (Africa, Missouri, USA). This is analogous to a text only index where extra weight is given to geographic references and synonyms expanded. The motivation behind this method is to maximise the recall of the geographic index.

Most referred to (MR). For each placename occurrence, classify as the matching location that is referred to most often. The motivation behind this method is to provide a base-line to see if using sophisticated methods significantly improves results.

C-Index (CI). A co-occurrence index is assigned to every triplet of adjacently occurring placenames. The c-index represents the confidence with which a triple can be disambiguated as the most likely occurring locations. Triplets are disambiguated in descending order of c-index and disambiguated locations are propagated to neighbouring triplets. The motivation of this method is to see if only adjacent placenames are needed for disambiguation and if propagation of disambiguated placenames improves accuracy.

Decision List (DL). In this method of disambiguation we infer a decision list from the corpus and disambiguate each placename independently using the rule with the greatest confidence. The algorithm was originally suggested by Rivest [10], the version of the algorithm we use is described in detail by Yarowsky [13] and is similar to the work done by Smith and Mann [11]. The motivation of this method is to see if first order co-occurrence is all that is necessary for placename disambiguation.

Support Vector Machine (SVM). In the final disambiguation method we approach placename disambiguation as a vector space classification problem. In this problem the placenames can be considered as objects to be classified and the possible locations as classification classes. The chosen features were the placenames that co-occur with the placename being classified. The scalar values of

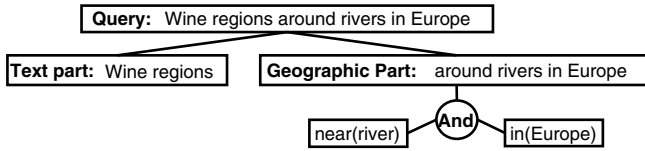


Fig. 2. Query Trees

the features are the inverse of their distance from the placename being classified, their sign is governed by whether they appear before or after the placename being classified. A separate feature space was built for each placename and linearly partitioned with a Support Vector Machine [6]. The motivation of this method is to see if multiple orders of co-occurrence can improve accuracy.

2.5 Query Engine

The *Query Engine* re-orders the results of the text queries produced by Lucene using the geographic queries.

The queries are manually split into a text component and a geographic component. The text query is handled normally by Lucene, the geographic query is manually split into a tree of conjunctions and disjunctions.

Executing a text query. Once the news articles have been indexed with Lucene, the query terms will be extracted in the same way as the document terms, a similarity measure is taken between the query's terms and all indexed documents. The similarity function is given by the following expression:

$$\text{score}(q, d) = \frac{\sum_{t \in q} \text{tf}_t(d) \text{idf}^2(t) \text{norm}(d)}{\sqrt{\sum_{t \in q} \text{tf}_t^2(d)}},$$

where $\text{tf}_t(d)$ is the t term frequency for the given document d (in our case is 0 or 1), $\text{idf}(t)$ is the inverse document frequency of term t , and $\text{norm}(d)$ is a normalisation constant given by the total number of terms in document d . See the [1] for details.

The query tree. The query trees are constructed by hand. The nodes of the tree are either conjunctions or disjunctions while the leaves of the tree are spatial-relation, location pairs (see Figure 2).

Executing a query. The documents that match both the geographic and the text query are returned first (as ranked by Lucene). This is followed by the documents that hit just the text query. The tail of the results is filled with random documents.

3 Experimental Runs and Results

We have executed 24 runs with the GeoCLEF 2006 data: all mono-lingual, manually constructed English queries on an English corpus. The queries constructed

with: the query topic and description (TD); the query topic, description and narrative (TDN); the text part containing no geographic entities (Tx); and text part containing geographic entities (GTx). We have four different query constructions: TD-Tx, TDN-Tx, TD-GTx and TDN-GTx.

The query constructions were tabulated against the six disambiguation methods. As far as was possible we attempted to add no world knowledge, the query trees produced resemble trees that could be generated with a query parser.

Our [non-base-line] runs appeared between the 25% and the 75% quartiles for mean average precision with most results around the median (results presented here were not included in the quartile calculations). We applied two sets of significance testing to our per query results. The Friedman test for multiple treatments of a series is a non-parametric test that can show differences across multiple treatments. The Wilcoxon signed-rank test is a non-parametric test that shows if two independent treatments have a significant difference and if there is a difference, which treatment is significantly better [5].

Mean Average Precision							Distribution	
	NoGeo	NoDis	MR	CI	DL	SVM	Worst	4%
TD-Tx	8.3%	17.1%	16.3%	16.6%	16.7%	16.3%	Q1	15.6%
TDN-Tx	8.4%	17.9%	19.4%	17.9%	19.4%	19.4%	Median	21.6%
TD-GTx	18.6%	22.9%	22.5%	21.3%	22.6%	22.5%	Q3	24.6%
TDN-GTx	21.7%	21.6%	22.2%	18.8%	22.2%	22.2%	Best	32.2%

The runs consisting of Title, Description and Narrative (TDN) statistically significantly out performed the respective runs consisting of only Title and Description (TD) when no geographic entities were contained in the text part of the query (Tx). The runs with geographic entities in the text part of the query (GTx) statistically significantly out performed the respective runs without geographic entities in the text part of the query (Tx). There was no statistical significance between the different geographic index runs (NoDis, MR, CI, DL and SVM). The runs using a geographic index were statically significantly better than the run without (NoGeo).

4 Conclusions and Future Work

We can conclude that the combination of geographic and text indexes generally improves geographic queries. To maximise MAP geographic phrases should be included when querying *both* the geographic index and the text index. This conclusion is consistent with previous experiments on the GeoCLEF data [4,8]. We can also conclude that our experiments show no significant difference in the MAP achieved from using any of our methods for generating a geographic index. The inclusion of the narrative information increased MAP only when there was no geographic information contained in the text part of the query or no geographic index was used. This is because the narrative of a query specifies the geographic phrase in greater detail mainly adding information already contained in the geographic index.

With respect to our objectives we can conclude that the co-occurrence model accuracy agrees with the previous experiments conducted in [9] and that co-occurrence models are a suitable method of placename disambiguation. By increasing the size & accuracy of the co-occurrence model, increasing the number of queries and improving how the different indexes are combined, we believe in future experiments the improvement produced by disambiguating placenames will increase.

Lucene was applied in the default configuration and the text part of the queries were not altered in any way. We plan to experiment with suitable query weights for Lucene and try alternative configurations of the index. Ultimately we would like to combine the geographic and text indexes so that they can be searched and applied simultaneously. We also plan to implement a query parser to allow the queries to automatically be parsed into query trees; this would require a level of natural language processing.

References

1. Apache Lucene Project (Accessed 01 November 2006) (2006), <http://lucene.apache.org/java/docs/>
2. Clough, P., Sanderson, M., Joho, H.: Extraction of semantic annotations from textual web pages. Technical report, University of Sheffield (2004)
3. Guttman, A.: R-Trees, A dynamic index structure for spatial searching. In: SIGMOD International Conference on Management of Data (1984)
4. Hauff, C., Trieschnigg, D., Rode, H.: University of Twente at geoCLEF 2006: geofiltered document retrieval. In: Working Notes for GeoCLEF (2006)
5. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: 16th Annual ACM SIGIR (1993)
6. Joachims, T.: Advances in Kernel Methods - Support Vector Learning (1999)
7. Kinzler, D.: Wikisense - Mining the Wiki. In: Wikimania '05 (2005)
8. Martins, B., Cardoso, N., Chaves, M., Andrade, L., Silva, M.: The university of lisbon at GeoCLEF 2006. In: Working Notes for GeoCLEF (2006)
9. Overell, S., Rüger, S.: Identifying and grounding descriptions of places. In: SIGIR Workshop on Geographic Information Retrieval (2006)
10. Rivest, R.: Learning decision lists. Machine Learning (1987)
11. Smith, D., Mann, G.: Bootstrapping toponym classifiers. In: HLT-NAACL Workshop on Analysis of Geographic References (2003)
12. Yarowsky, D.: One sense per collocation. In: ARPA HLT Workshop (1993)
13. Yarowsky, D.: Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In: 32nd Annual Meeting of the ACL (1994)
14. Zhu, J., Uren, V., Motta, E.: ESpotter: Adaptive named entity recognition for web browsing. In: Professional Knowledge Management Conference (2005)

NICTA I2D2 Group at GeoCLEF 2006

Yi Li, Nicola Stokes, Lawrence Cavedon, and Alistair Moffat

National ICT Australia, Victoria Research Laboratory*
Department of Computer Science and Software Engineering
The University of Melbourne, Victoria 3010, Australia
{yli8,nstokes,lcavedon,alistair}@csse.unimelb.edu.au

Abstract. We report on the experiments undertaken by the NICTA I2D2 Group as part of GeoCLEF 2006, as well as post-GeoCLEF evaluations and improvements to the submitted system. In particular, we used techniques to assign probabilistic likelihoods to geographic candidates for each identified geo-term, and a probabilistic IR engine. A *normalisation* process that adjusts term weights, so as to prevent expanded geo-terms from overwhelming non-geo terms, is shown to be crucial.

1 Introduction

I2D2 (Interactive Information Discovery and Delivery) is a project being undertaken at National ICT Australia (NICTA), with the goal of enhancing user interaction with information hidden in large document collections. A specific focus of I2D2 is in detecting geographically salient relationships, with Geographic Information Retrieval (GIR) being an important challenge.

The system used in the I2D2 2006 submission to GeoCLEF is built on top of the Zettair [1] IR engine, extending the base engine with a probabilistic retrieval technique to handle ambiguous geographic references. The documents are pre-processed with Language Technology components to identify and resolve geo references: we use the LingPipe *named entity recognition and classification* system; and a *toponym resolution* component. The toponym resolution component assigns probabilistic likelihoods to potential candidate locations, found in a gazetteer, for geographic terms identified by LingPipe.

This paper describes the system architecture and a number of experiments performed for the purpose of GeoCLEF, as well as evaluating different approaches to geospatial-retrieval. In particular, we experimented with both *document expansion* and *query expansion*, that is, replacing geographic terms in documents or queries with a list of related terms, as described below. To combat the drop in precision resulting from geographic expansion, as noted by GeoCLEF 2005 participants (for example, [2]), we implemented a *normalization step* to ensure that the added location names do not overwhelm other terms in the topic. The submitted runs used an early version of this normalization; a subsequent

* NICTA is funded by the Australian Government's "Backing Australia's Ability" initiative, in part through the Australian Research Council.

refined version saw a slight increase in overall MAP over the non-GIR baseline run. During these post-submission experiments we also noticed that our linguistic annotation was having an unpredictable effect on document rankings, due to the increase in average document length (in bytes) which is used by the Okapi BM-25 ranking method [3]. This problem is also address in this paper, and is described in Section 3.

Overall, our baseline for GeoCLEF 2006 topics seemed low (MAP score of 0.2312); adding the geographic retrieval techniques increased overall MAP by 1.86% (document expansion) and 3.11% (query expansion) over the baseline. Variation in performance was seen across topics; this is discussed in Section 4.

2 System Description

There are four steps involved in our probabilistic geospatial information retrieval (GIR) process: *named entity recognition and classification* (NERC); *probabilistic toponym resolution* (TR); *geo-spatial indexing*; and *retrieval*. For maximum flexibility and easy re-configuration, we used the UIMA¹ document processing architecture to augment documents with extra annotations and to perform the indexing.

We used a named entity recognition and classification system to differentiate between references to the names of places (which we are interested in), and the names of people and organizations (which we are not). A surprising number of everyday nouns and proper nouns are also geographic entities, for example, the town “Money” in Mississippi. Errors in this part of the pipeline can have a significant effect on the accuracy of the disambiguation process. Our system uses the LingPipe open-source NERC system, which employs a Hidden Markov model trained on a collection of news articles (<http://www.alias-i.com/lingpipe/>). For further GIR system details, see [4].

Toponym Resolution

Toponym resolution (TR) is the task of assigning a location to each place name identified by the named entity recognizer. Many place names are ambiguous; context surrounding a place name in the text can be used to determine the correct candidate location. Our approach to TR assigns probability scores to each location candidate of a toponym based on the occurrence of hierarchical associations between place names in the text. Hierarchical associations and location candidates pertaining to a particular geographical reference can be found in a gazetteer resource. For this purpose, we used the Getty Thesaurus, available from <http://www.getty.edu/vow/TGNServlet>.

Probabilities are allocated to candidate locations based on a five-level normalization of the gazetteer. For example, a candidate that is classified as a *continent* or *nation* receives a significant probability, while a candidate that is classified as

¹ <http://www.research.ibm.com/UIMA/>

an *inhabited place* (which includes cities) initially receives a much smaller probability, and so on. Initial probability assignments are then adjusted based on a variety of evidence, such as: *local contextual information*, for example, geo-terms occurring in close proximity mutually disambiguate each other, in particular, city-state pairs; *population information*, when available; *specified trigger words* such as “County” or “River”; and *global contextual information*, such as occurrences in the document of “country” or “state” type geo-terms that are gazetteer ancestors to the candidate. Final probability assignments are then normalized across the complete set of possible candidates for each geo-term.

We used a hand annotated subset of the GeoCLEF corpus to determine the performance of the Named Entity Classification system, and our toponym disambiguation algorithm. This annotated corpus consisted of 302 tagged GeoCLEF news articles; a total of 2261 tagged locations. The overall precision of LingPipe on this dataset was 46.88% precision and 69.77% recall. With respect to disambiguation accuracy our system achieved an accuracy of 71.74%. Using additional geographical evidence from Wikipedia we were able to increase the accuracy to 82.14%. A detailed description of this experiment, as well as further detail on the impact of NLP errors, is presented in [5].

Probabilistic Geographical IR

Our Geographical Information Retrieval system involves an extension of Zettair [1], to which we add spatial-term indexing. Hierarchically expanded geo-terms (in each case a concatenated string consisting of a candidate and its ancestors in the gazetteer) are added to an index. Geo-tagged queries can then be processed by matching geo-terms in the query to geo-terms in the spatial index.

The system supports both *document expansion* and *query expansion* techniques for matching the location in a query to all its gazetteer children and nearby locations. Document expansion (or *redundant indexing*) involves adding spatial terms to the index for each of a geo-term’s ancestors in the gazetteer hierarchy. Query expansion involves expanding terms in the query. This technique allows more flexible weighting schemes, whereby different weights can be assigned to documents which are more relevant at different hierarchical levels or spatial distances.

A geo-term may be expanded either *upwards* or *downwards*. Downward expansion extends the influence of a geo-term to some or all of its descendants in the gazetteer hierarchy to encompass locations that are part of, or subregions of, the specified location. Upward expansion expands the influence of a geo-term to some or all of its ancestors, and then possibly downward to siblings of these nodes. For example, downward expansion was used for geo-terms preceded by an “in” spatial relation, while upward expansion was used for “close/near” relations.

After expansion, weights are assigned to all expanded geo-terms, reflecting their estimated similarities to the source query geo-term. We used *hierarchical distance* for downward expansion and *spatial distance* for upward expansion. Finally, the *a priori* Okapi BM-25 approach [3] (as implemented in Zettair) is

used to calculate the sum of scores for the query. We apply a *normalization step* to obtain a single score for each location concept by combining the similarity scores of its geo-term, text term, and expanded geo-terms. Without this step, irrelevant documents that contain many of the expanded geo-terms in the query will be incorrectly favored. The contribution of the (potentially numerous) geo-terms added to an expanded query might then overwhelm the contribution of the non-geo terms in the topic. Our ranking algorithm works as follows.

In a query, all query terms are divided into two types: *concept terms* t_c and *location terms* (annotated with reference to gazetteers) t_l . In a query example “wine Australia”, the concept term is “wine” and the location term is “Australia”. The final score is contributed to by both concept and locations terms:

$$\text{sim}(Q, D_d) = \text{sim}_c(Q, D_d) + \text{sim}_l(Q, D_d) \quad (1)$$

where $\text{sim}_c(Q, D_d)$ is the concept similarity score and $\text{sim}_l(Q, D_d)$ is the location similarity score. The concept similarity score is calculated using the same calculation as Okapi. The location similarity score can then be denoted as:

$$\begin{aligned} \text{sim}_l(Q, D_d) &= \sum_{t \in Q_l} \text{sim}_t(Q, D_d) \\ &= \sum_{t \in Q_l} \text{Norm}_t \left(\text{sim}_{\text{text}}(Q, D_d), \right. \\ &\quad \left. \text{sim}_{\text{geoI}}(Q, D_d), \dots, \text{sim}_{\text{geoT}}(Q, D_d) \right) \end{aligned} \quad (2)$$

where Q_l is the aggregation of all location terms in the query, T is the number of all corresponding geo-terms (including expanded geo-terms) of a location $t \in Q_l$, and $\text{Norm}_t()$ is a normalization function which normalizes the similarity scores of a location t 's textual terms, geo-term and its expanded geo-terms.

To define the normalization function, assume that we have T similarity scores from terms belonging to the same location (text, geo-term and expanded geo-terms), and after sorting them into descending order, they are: $\text{sim}_1, \dots, \text{sim}_T$. We use a geometric progression to compute the final normalization score:

$$\text{Norm}(\text{sim}_1, \dots, \text{sim}_T) = \text{sim}_1 + \frac{\text{sim}_2}{a} + \dots + \frac{\text{sim}_T}{a^{T-1}} \quad (a > 1) \quad (3)$$

3 Experimental Results

All of our GeoCLEF 2006 submitted runs were based on the Zettair system, some using baseline Zettair and others using the probabilistic IR techniques described in the previous section. The runs submitted were:

1. **MuTdTxt**: Baseline Zettair system run on unexpanded queries formed from topic *title* and *description* only. We take this to be our baseline.
2. **MuTdnTxt**: Baseline Zettair system run on unexpanded queries formed from topic *title* and *description* and location words from the *narrative* field.

3. **MuTdRedn**: Baseline Zettair system run on queries formed from topic *title* and *description*. Documents are automatically toponym-resolved and expanded with related geo-terms (as described in the previous section). The query (*title* and *description*) is automatically annotated, geo-terms are disambiguated, but geo-terms are *not* further expanded with related geo-terms.
4. **MuTdQexpPrb**: Probabilistic version of Zettair. Documents are automatically annotated and disambiguated but are *not* expanded with related geo-terms. Query (*title* and *description*) is automatically resolved for geo-terms, and geo-terms are expanded with related geo-terms. This is our most complete Geographic IR configuration.
5. **MuTdManQexpGeo**: Baseline Zettair using purely text-based retrieval, but for which the query is *manually* expanded (with related text place-names).

Table 1 shows overall mean average precision (MAP) scores for runs submitted to GeoCLEF. These scores are significantly lower than the MAP score obtained by the baseline system (**MuTdTxt**) run over GeoCLEF 2005 topics: 0.3539.

Table 1. MAP scores for each submitted run over all 25 topics

Run	MuTdTxt	MuTdnTxt	MuTdRedn	MuTdQexpPrb	MuTdManQexpGeo
MAP	0.2312	0.2444	0.2341	0.2218	0.2400
% Δ		+5.71%	+1.25%	-4.07%	+3.81%

Subsequent to submitting to GeoCLEF 2006, we made some improvements to the normalization step described in Section 2. Further, we discovered, after annotation, that documents have different changes in length ratio, depending on the density of the spatial references they contain. This change has an impact on the final ranking of these documents. To evaluate the impact of different length normalization strategies, we performed some further experiments. First, the unannotated collection was indexed. The average document length was 3532.7 bytes. We then indexed an annotated version of the collection, in which the average document length was 4037.8 bytes. Finally, we executed 25 topics against these two indexes using only text terms, and found large variations in the average precision value differences (standard deviation of 15.8%). To counteract the effect of byte length variations, all the annotated documents should use the same document byte length as their baseline unannotated documents when indexed. We used this technique to index the same annotated collection. From Figure 1 we can see that for most of the topics their MAP is not changed at all, the variations in the average precision value differences obtained is very small with a stadand deviation of only 0.71%.

We re-ran our two geographic runs again using both improved normalization and baselined document length. The results are provided in Table 2.

Figure 2 displays the average precision scores for each topic and for each run, after including the newer, improved normalization step. There is a high degree of variance in the performance obtained across the set of queries. To discover

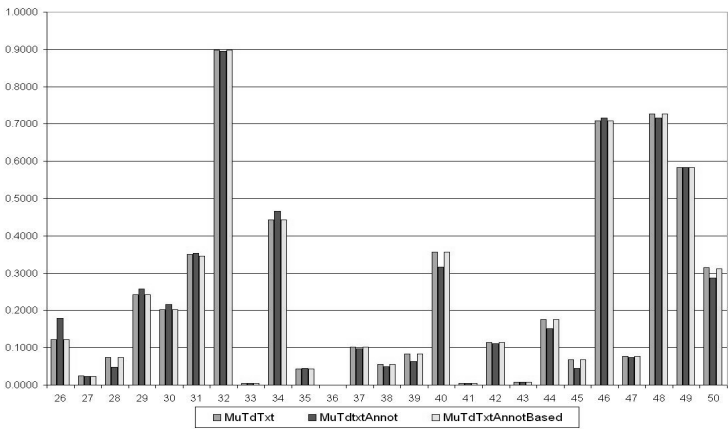


Fig. 1. Average Precision per topic, for each run, of three different text-only runs. TTxT: unannotated documents; TTxTAnnot: annotated documents with different document length; TTxTAnnotBased: annotated documents with baselined document length.

Table 2. MAP scores over all 25 topics, using improved normalization and baselined document length

Run	MuTdTtxt	MuTdnTxt	MuTdRedn	MuTdQexpPrb	MuTdManQexpGeo
MAP	0.2312	0.2444	0.2355	0.2384	0.2400
%Δ		+5.71%	+1.86%	+3.11%	+3.81%

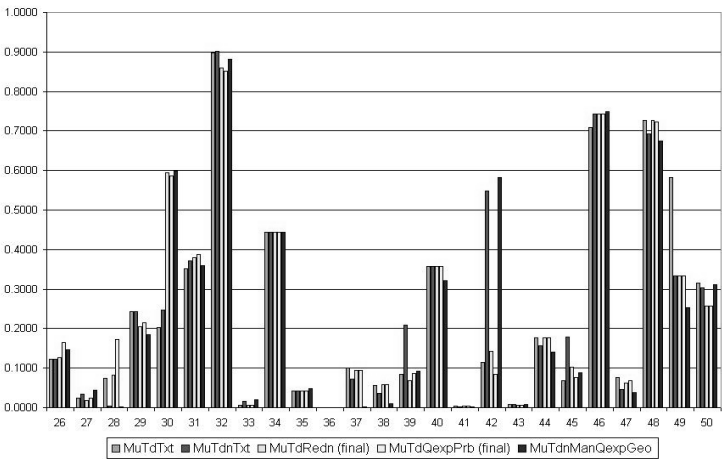


Fig. 2. Average Precision per topic, for each run, using improved normalization and baselined document length

the performance of our query normalization, we also performed a further run without normalization. Its overall MAP is 0.1994, which is a significant decline.

We also performed a non-probabilistic run whose overall MAP is 0.2388, indicating little difference between probabilistic and non-probabilistic techniques.

4 Analysis and Conclusions

One of the underlying assumptions of our current GIR architecture is that every query that makes reference to a geographical entity requires geospatial expansion. From a detailed analysis of our results we observed that this is not always the case. For example, topic 28 (*Snowstorms in North America*) benefits more from concept expansion (“snow”, “storm”) than geo-term expansion. The frequent occurrence of queries of this nature in the GeoCLEF topics may explain the high performance of GeoCLEF systems that disregard the geographical nature of the queries and expand instead through relevance feedback methods [2]. It may also explain why our manual expansion run did not significantly outperform the baseline run, as the human annotator added geo-terms only.

However even when queries are suitable candidates for geospatial expansion, the query expansion methods describe in this paper often do not perform as expected. One reason for this is that our gazetteer lacks some vital information needed for geospatial expansion. For example, topics that mention politically defunct locations, such as “former Yugoslavia” and “the eastern Bloc”, are not listed by the Getty thesaurus. In addition, locations that fall under the “general region” category in Getty, such as “Southeast Asia”, “the Ruhr area” and “the Middle East”, cannot be expanded as neither their related children nor their associated long/lat coordinates are listed in the gazetteer. This explains the poor to average performance of our geo-runs on topics 33, 34, 35, 37, 38, and 44.

Another assumption of our GIR architecture is that GeoCLEF queries that will benefit from geospatial expansion only require the additional concepts provided by a standard gazetteer; that is, neighboring locations and synonyms. Some GeoCLEF queries require adjective-form expansion of their geo-terms; this is particularly important when a document mentions the concept of the query (for example, “diamonds” in topic 29) but alludes to the location of the event through the use of a geographical adjective (for example, “a South African-based organization”). Examples such as these indicate that metonymic references to locations do not necessarily reduce the performance of GIR systems, as was implied in [6]. The adjective form of place names can be captured from a thesaural resource, such as WordNet.

In addition, some geospatial queries require entity type expansion that is constrained by geographic boundaries. For example, in topic 43 (*Scientific research in New England universities*), ignoring the syntactic relationship between the location “New England” and the entity “universities”, will ensure that documents that mention “Harvard” will appear less relevant, since Harvard isn’t (by coincidence) a state, county, town or suburb in New England. Information such as this may only be found using an additional knowledge source such as Wikipedia.

Despite these drawbacks, expansion of the query has had a positive effect on some topics, with the normalization step seeming to have alleviated the query overloading problem. Topic 26 (*Wine regions around rivers in Europe*) sees an increase over both baseline and the manually-annotated-query runs. Improvement over baseline was also seen with queries involving spatial relations such as “near” (for example, topic 30, *Car bombing near Madrid*); and with geographic specialization, such as *Northern* (for example, topic 42, *Regional elections in Northern Germany*). Note, however, that our handling of spatial relations did not extend to specific distance specifications, which may have contributed to a slight drop in precision from baseline for topic 27 (*Cities within 100km of Frankfurt*).

Missed recognitions and misclassifications by LingPipe, and incorrectly disambiguated locations in the queries and collection, will also have compromised the performance of our GIR runs. In additional experiments, not discussed in this paper, we have found that NERC systems such as LingPipe are significantly underperforming on GeoCLEF relative to the performance scores reported at evaluation forums such as MUC;² re-training of off-the-shelf systems for GeoCLEF will significantly reduce such errors. However, some NERC errors are more critical than others; maximizing NERC recall rather than precision will improve the quality of the annotated data presented to the GIR system, as the toponym resolution step filters out much of the NERC misclassified named entities. Further discussion of this work can be found in [5].

References

1. Zettair: The Zettair search engine, <http://www.seg.rmit.edu.au/zettair/index.php>
2. Gey, F., Petras, V.: Berkeley2 at GeoCLEF: Cross-language geographic information retrieval of German and English documents. In: GeoCLEF 2005 Working Notes, Vienna (2005), http://www.clef-campaign.org/2005/working_notes/
3. Walker, S., Robertson, S., Boughanem, M., Jones, G., Sparck Jones, K.: Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering and QSDR. In: Proc. Sixth Text Retrieval Conference (TREC 6), Gaithersburg, Maryland (November 1997)
4. Li, Y., Moffat, A., Stokes, N., Cavedon, L.: Exploring probabilistic toponym resolution for geographical information retrieval. In: SIGIR Workshop on Geographical Information Retrieval, Seattle (2006)
5. Stokes, N., Li, Y., Moffat, A., Rong, J.: An empirical study of the effects of NLP components on geographic IR performance. International Journal of Geographical Information Science, special issue on Geographical Information Retrieval
6. Leveling, J., Hartrumpf, S.: On metonymy recognition for geographic ir. In: SIGIR Workshop on Geographical Information Retrieval, Seattle (2006)

² http://www-nlpir.nist.gov/related_projects/muc/

Blind Relevance Feedback and Named Entity Based Query Expansion for Geographic Retrieval at GeoCLEF 2006

Kerstin Bischoff, Thomas Mandl, and Christa Womser-Hacker

University of Hildesheim, Information Science
Marienburger Platz 22, D-31141 Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. In its participation at GeoCLEF 2006, the University of Hildesheim focused on the monolingual German and English and the bilingual German $\leftrightarrow$ English tasks. Based on the results of GeoCLEF 2005, the weighting of and the expansion with geographic named entities (NE) within a Boolean retrieval approach was examined. Because the best results 2005 were achieved with Blind Relevance Feedback (BRF) in which NEs seemed to play a crucial role, the effects of adding particular geographic NEs within the BRF are explored. The paper presents a description of the system design, the submitted runs and results. A first analysis of unofficial post experiments indicates that geographic NEs can improve BRF and supports prior findings that the geographical expansion within a Boolean retrieval approach does not necessarily lead to better results – as it is often assumed.

1 Introduction

Many queries posted to search services are of geographic nature, i.e. the information searched for is restricted to a certain geographic region or place. The development of Geographic Information Systems (GIS) for using structured spatial data e.g. in urban planning or for new location based services has long attracted much interest. Geographic Information Retrieval (GIR) is a comparatively new field of research. Having participated in other CLEF tracks before, the University of Hildesheim experimented in GeoCLEF 2006 in order to evaluate different approaches for the retrieval of geo-referenced information from unstructured data.

In GeoCLEF 2005, various approaches ranging from basic information retrieval (IR) techniques to elaborated spatial indexing and retrieval methods were used [6]. The most successful runs were based on a fine-tuned BRF [7]. Gey and Petras found that improvement through BRF seemed highly related to adding proper names as “good” terms to the original query [7]. Thus, the expansion by primarily geographic named entities (NE) might further improve retrieval quality. The experiments reported in this paper focus on the role of geographic NEs within the process of BRF.

In contrast, the results of last year’s GeoCLEF track showed worse retrieval performance after manual or automatic expansion of geographic NEs to include their finer-grained sub regions resp. more specific location names; even combined with a

Boolean approach the results were mixed [5,7,9]. Therefore the task for 2006 was especially designed to explore the usefulness of additional geographic information (mostly country names) provided in the narratives of the topics. Hence, we experimented with the recognition, weighting and expansion of such geographic NEs (not) using Boolean conjunction for German and English monolingual and German $\leftrightarrow$ English bilingual retrieval.

2 CLEF Retrieval Experiments with the MIMOR Approach

A system applied to ad-hoc retrieval in previous CLEF campaigns [8] was augmented for experiments with (geographic) NEs in GIR. It uses Lucene's¹ technology for indexing and searching based both on similarity as well as the Boolean model. For bilingual runs topics were first translated by combining three online translation services². By merging multiple translation services the influence of wrong translations should be minimized. The University of Neuchatel's stopword lists³ for English and German were used and augmented with common CLEF words for topic formulation. Morphological analysis was done by the Lucene stemmer for German and the Snowball stemmer for English.

For Named Entity Recognition (NER), we employed the open source machine learning tool LingPipe⁴, which identifies named entities and classifies them into the categories *Person*, *Organization*, *Location* and *Miscellaneous* according to a trained model. For English, the model provided by LingPipe was used. It is trained on newspaper data. For German, a model trained on an annotated corpus of German newspaper articles (*Frankfurter Rundschau*) [10] was used. NER was applied for query processing for weighting NEs and to generate Boolean queries of the type *concept AND geographical reference*. NER was also applied to the document collection for building an index with separate fields for the NE categories.

According to Gey and Petras [7], NEs seem to be a crucial factor for successful BRF. Since the "good" terms to be added to the original query were mostly proper nouns, the addition of primarily geographic names during the BRF process should further improve retrieval quality. This might be particularly promising for imprecise regions like *Northern Germany* or other geographic names that are not to be found in a geographic thesaurus (*gazetteer*). Comprehensive gazetteers are often not publicly available. Adequate heuristics need to be established on which information to extract in order to effectively expand a query. Thus, looking for frequent co-occurrences of the geographic NE of a query with other geographic NEs within the top-ranked documents may provide hints for the appropriate kind of information and enable the inclusion of more specific location names without the help of a gazetteer (e.g. *Ruhrgebiet – Bochum, Gelsenkirchen*).

The following steps in query processing and retrieval were carried out in our system (optional steps in parentheses): topic parsing, (topic translation), (NER and

¹ <http://lucene.apache.org/java/>

² <http://babelfish.altavista.com>, <http://www.linguatec.de>, <http://www.freetranslation.com>

³ <http://www.unine.ch/info/clef/>

⁴ <http://www.alias-i.com/lingpipe>

weighting), stopword removal, stemming, Boolean or ranked (Boolean) query, (BRF with or without NE weighting; Boolean or ranked expansion)

3 Submitted Runs

After experimentation with the GeoCLEF 2005 data, runs differing in several parameters were submitted. For monolingual English, the base runs applied no additional steps, no NER and no BRF. To test the effect of geographic expansion through BRF, two runs were submitted, in which (geographic) NEs were recognized and weighted as well as primarily extracted from the top-ranked five documents in the BRF process and added to the geographic clause of a Boolean query. In a fifth run, the geographic entities from the BRF were added to a similarity based retrieval model. For the monolingual German task, we submitted runs with similar parameters as for English. For comparison, two runs with “traditional” BRF not highlighting any NEs at all were produced. Since German base runs without BRF performed poorly with respect to the topics of GeoCLEF 2005, no such official runs were submitted.

For the bilingual tasks, the parameters were identical to those runs for monolingual retrieval. Run descriptions and results are shown in table 1 and table 2.

Table 1. Results monolingual runs

Run Identifier	Lang.	Fields	NEs weighted	BRF			Query	MAP
				Docs.	Terms	weighted		
HIGeoenenrun1	Engl	TD	–				OR	16.76
HIGeoenenrun1n	Engl	TDN	–				OR	17.47
HIGeoenenrun2	Engl	TD	w	5	25	GeoNEs and NEs	AND	11.66
HIGeoenenrun2n	Engl	TDN	w	5	25	GeoNEs and NEs	AND	12.13
HIGeoenenrun3	Engl	TD	–	5	20	GeoNEs	OR	18.75
HIGeodederun4	Germ	TD	–	5	25		OR	15.58
HIGeodederun4n	Germ	TDN	–	5	25		OR	16.01
HIGeodederun6	Germ	TD	w	5	25	GeoNEs and NEs	AND	12.14
HIGeodederun6n	Germ	TDN	w	5	25	GeoNEs and NEs	AND	11.34

The results reached the average performance of all participants for monolingual and bilingual tasks. A first look at the MAPs for single topics revealed that performance was best for five topics, which were more of ad-hoc style (GC 29, 30, 32, 46, 48). The relatively low scores may indicate that geographic IR is a special task and outline the importance of evaluation methods. Though geographic BRF adding 25 terms from five documents combined with the Boolean approach worsened performance for all tasks compared to the respective base runs, geographic BRF alone seemed to improve retrieval quality at least slightly. Query expansion by additional geographic information from the narrative field led neither to substantial improvement nor deterioration of performance. The only notable difference was found for bilingual German → English base runs, where improvement was mainly due to a higher

Table 2. Results bilingual runs

Run Identifier	Lang.	Fields	NEs weighted	BRF			Query	MAP
				Docs.	Terms	weighted		
HIGeodeenrun11	De → En	TD	–				OR	15.04
HIGeodeenrun11n	De → En	TDN	–	–			OR	19.03
HIGeodeenrun13	De → En	TD	w	5	25	GeoNEs, NEs	AND	14.56
HIGeodeenrun13n	De → En	TDN	w	5	25	GeoNEs, Nes	AND	15.65
HIGeodeenrun12	De → En	TD	–	5	20	GeoNEs	OR	16.03
HIGeodeenrun21	En → De	TD	–	5	25		OR	11.86
HIGeodeenrun21n	En → De	TDN	–	5	25		OR	13.15
HIGeodeenrun22	En → De	TD	w	5	25	GeoNEs, Nes	AND	09.69
HIGeodeenrun22n	En → De	TDN	w	5	25	GeoNEs, Nes	AND	10.46

accuracy for topic GC 38 (100% with using the narrative), a topic which only had one relevant document in the (pooled) collection. Bilingual retrieval performance relies on translation quality. Measuring translation quality is difficult, but merging the three translation services seems to work well considering particularly wrong translations. Especially translating English into German resulted in many errors, while the opposite direction did produce much better translations ambiguity in the target language (e.g. *banks of European rivers* → *Bänken von europäischen Flüssen* or confusion between parts of speech e.g. *car bombings near Madrid* → *Autombomben nähern sich Madrid*) should have a significant impact on retrieval quality. Problems decomposing geographic entities during translation also occur (e.g. *new narrow country* instead of *New England*, the *chew case mountains* instead of *Caucasus Mountains*). Transliterations and the word by word translation of compound geographic NEs can also be found (e.g. *Middle East/Naher Osten*; *Vila Real* → *Vila actually*; *Oman* → *grandmas*).

Though it would be ideal to have a comprehensive gazetteer including name variants for all possible languages, merging could almost always assure that at least one correct translation was contributed to the automatically translated query. Critical for all translation services were geographic entities like *Westjordanland* (*Westbank*), *Berg-Karabach/ Nagorno-Karabakh*, the *Sea of Japan/ das Japanische Meer*, *Mediterranean sea* (→ *Mittelmeermeer*) and *Ruhr area*.

Combining multiple translation services could on the other hand considerably improve NER for German. The evaluation of the NER revealed an unsatisfying rate of approximately 45% correct recognition and classification of geographic entities (incl. *Warsaw Pact* and *the Eastern Bloc*) for German monolingual runs. After the translation the rate improved by 7 % (52 %). Considerable improvement can be seen within the narrative of topic GC 37 (*Middle East*) where much better NER results were achieved. Listings of geographic entities within the narratives was a major difficulty for NER in German except for the last topic GC 50. Since the formulation of some topics did not vary between title and description as far as grammatical embedding of geographic NEs is concerned (e.g. GC 31, 33), all of them were missed. For English, an effect of translation was not observed. The NER rate

decreased from 72% to 70%. Error during translation also introduced some false positives, which would later be weighted high in the query. Automatic translations of very low quality did not lead to false positives, but the NER module made similar mistakes or missed the same geographic entities. The recognition of the correct boundaries of compound names proved to be difficult (e.g. *the Sea of Japan*, *the East Sea of Korea* or *the Warsaw Pact*).

4 Discussion of Post Submission Runs

Since the resulting MAPs of our submitted runs, whose parameters had been tuned to fit the GeoCLEF Data of 2005, did not differ substantially, additional post experiments were run trying to isolate the single effects of BRF, NER and weighting and Boolean conjunction. A German base run without any BRF and weighting was not submitted due to an encoding problem. In post experiments, such a German base run yielded 25.73% MAP using title and description. Table 3 shows MAP-scores of similarity based monolingual experiments.

GeoBRF does lead to considerable better performance for German. Whereas BRF not highlighting (geographic) NEs achieved lower MAPs than the base runs for both languages, GeoBRF could significantly improve the average precision scores for German. The best BRF parameters were 30 terms from ten top-ranked documents, which were weighted about half as strong as original query terms. In combination with weighting geographical entities high and other NEs moderately higher than the remaining terms performance further improved over the base run. The influence of non-geographic NEs in weighting the query before BRF can be neglected (maximum 1% change in MAP) probably because of the nature of the GeoCLEF topics. The best GeoBRF, however, did not disregard other NEs, but boosted them about a third as much as geographic ones. For English, GeoBRF did not or only marginally improve results. That could be related to the fact that some topics have few relevant documents. Half of the topics have less than ten relevant documents in the collection and only some had more than 20 relevant documents. That might have reduced the pool of “good” documents to take extract expansion terms from.

For German monolingual retrieval, GeoBRF added terms like *Regensburg*, *Straubing*, *Wachau* for cities along *the Danube* or *the Rhine* (GC 50), *Lower Saxony*, *Bremen*, *Hamburg* for *Northern Germany* (GC 42) or *Mühlheim* and *Gelsenkirchen* for *the Ruhr* (GC 33).

It might be promising to analyse in more detail, for which topics the strategy fits best. It seems that using the additional information provided by the topic narratives aggravates the problem of misleading geographic NEs. Considering the improvement in performance by GeoBRF in contrast to “traditional” BRF it would be worth integrating external resources like a gazetteer to only extract names that are within a certain region (e.g. belong to the same branch within a hierarchy). Methods for finding the geographic scope and focus of documents [1] could further help selecting the geographically best documents among the top-ranked documents.

Table 3. Monolingual German and English runs with (Geo)BRF

Fields	(Geo)NEs	BRF			German MAP	English MAP
		Docs	Terms	GeoBRF		
TD	–	–	–	–	25.73	18.11
TDN	–	–	–	–	23.43	16.27
TD	–	5	25	–	19.72	15.71
TDN	–	5	25	–	22.28	15.47
TD	–	10	20	–	21.82	15.74
TDN	–	10	20	–	21.05	16.84
TD	–	10	30	–	19.25	14.66
TDN	–	10	30	–	21.60	16.41
TD	–	5	25	yes	26.43	19.00
TDN	–	5	25	yes	25.71	18.12
TD	–	10	20	yes	27.36	15.69
TDN	–	10	20	yes	27.04	14.22
TD	–	10	30	yes	30.34	17.44
TDN	–	10	30	yes	28.65	15.53
TD	weighted	–	–	–	27.67	20.38
TDN	weighted	–	–	–	25.65	18.42
TD	weighted	5	25	–	23.59	19.03
TDN	weighted	5	25	–	23.33	18.30
TD	weighted	10	20	–	24.72	18.65
TDN	weighted	10	20	–	24.70	17.70
TD	weighted	10	30	–	24.98	17.88
TDN	weighted	10	30	–	24.03	18.04
TD	weighted	5	25	yes	29.54	18.65
TDN	weighted	5	25	yes	25.65	18.20
TD	weighted	10	20	yes	29.07	14.70
TDN	weighted	10	20	yes	27.90	14.83
TD	weighted	10	30	yes	31.20	17.84
TDN	weighted	10	30	yes	28.72	17.50

In general the MAPs of the runs with and without narratives were comparable. Since geographic expansion by GeoBRF worked well, this may indicate that the narratives do perhaps not contain the right kind of additional geographic information for most topics. To evaluate the expansion through narrative information in combination with Boolean retrieval, experiments varying NER, weighting, BRF and GeoBRF were run. For German, the results were poor. BRF with 20 terms from ten documents did better than any GeoBRF. Therefore, only monolingual English runs are shown in table 4.

The best English run did not apply NER, weighting or (Geo)BRF. It was superior to the non Boolean base run. Weighting of geographic NEs decreased quality, whereas other NEs had no influence. Regarding BRF the number of documents seems crucial to success. With fewer documents and without the narrative GeoBRF worsened performance. Thus, whereas without BRF or GeoBRF using the narrative information decreased average precisions for English (and

Table 4. (Geo)BRF for English monolingual runs using Boolean operators

Fields	NEs	BRF			English MAP
		Documents	Terms	GeoBRF	
TD	–	–	–	–	21.23
TDN	–	–	–	–	17.60
TD	weighted	–	–	–	21.23
TDN	weighted	–	–	–	19.41
TD	weighted	10	20	–	18.99
TDN	weighted	10	20	–	20.05
TD	–	5	25	–	18.09
TDN	–	5	25	–	19.49
TD	–	10	30	–	18.95
TDN	–	10	30	–	20.18
TD	–	10	20	yes	18.04
TDN	–	10	20	yes	20.91
TD	–	5	25	yes	14.08
TDN	–	5	25	yes	14.91
TD	–	10	30	yes	13.49
TDN	–	10	30	yes	18.56

German) by up to 7% MAP, BRF and GeoBRF reduced this negative impact substantially and even produced higher MAPs.

Moreover, we experimented with first running a ranking query and subsequently adding geographic NEs from the BRF as a Boolean clause. For both German and English, this led to worse results compared to the base runs. Nevertheless, this technique could improve MAP by 2% for English and by 3% for German by weighting NEs in the original OR-query and then expanding via GeoBRF with the GeoNEs combined via AND. The optimal number of (geographic) terms to add in the BRF was 25 from the five top-ranked documents for English and 30 terms from 15 documents weighted about half as much as the original query. Expanding with narrative information led to poor results. The impact of additional geographic information remains to be fully understood. Particularly different kinds of geographic information to use for expansion must be tested. Experiments with manual expansion of German queries support the assumption that the narrative information may not be adequate for the German collection.

5 Conclusion and Outlook

Query expansion by adding geographic NEs from the top-ranked documents within BRF has shown to substantially improve performance. The Boolean approach to GIR is not yet fully understood. Current work is focused on improving the NER by the fusion of machine learning with alternative approaches like gazetteer-look up and handcrafted rules. The integration of a gazetteer for automatic query expansion is planned. Priority is given to the exploration of heuristics for such an expansion.

To pursue the idea of geographic BRF, techniques for disambiguation and identification of the geographical scope of a document [1] should be integrated. Such

approaches would also be crucial for using Wikipedia as an additional external resource, by which geographic references not captured in gazetteers should be expandable. Expansion with synonyms e.g. via the linguistic service *Wortschatz Leipzig*⁵ for German is also envisioned. A detailed analysis of system performance on the individual topics however has to show the feasibility of a text based approach without any Geo-Coding and -matching. Some topics of this year would seem to demand elaborated spatial methods or techniques of natural language processing (NLP) to identify place types e.g. cities and negations (e.g. *except the republics of the former USSR*). However, some of these topics might not mirror realistic user needs.

References

1. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-Where: Geotagging Web content. In: Proc. 27th Annual Intl ACM SIGIR Conf. 2004, Sheffield, UK, pp. 273–280. ACM, New York (2004)
2. Chaves, M., Martins, B., Silva, M.: Challenges and Resources for Evaluating Geographical IR. In: Proc 2nd Intl Workshop on Geographic Information Retrieval, CKIM 2005, Bremen, Germany, pp. 65–69 (2005)
3. Clough, P.: Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. In: Proc 2nd Intl Workshop on Geographic Information Retrieval, CKIM 2005, Bremen, Germany, pp. 25–30 (2005)
4. Purves, R., Clough, P., Joho, H.: Identifying imprecise regions for geographic information retrieval using the web. In: Proc GIS Research UK 13th Annual Conference, Glasgow, UK, pp. 313–318 (2005)
5. Ferrés, D., Ageno, A., Rodríguez, H.: The GeoTALP-IR System at GeoCLEF-2005: Experiments Using a QA-based IR System, Linguistic Analysis, and a Geographical Thesaurus. In: Working Notes 6th Workshop Cross-Language Evaluation Forum, CLEF, Vienna (2005)
6. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P.: GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF, Vienna, Austria, September 2005 (2005)
7. Gey, F., Petras, V.: Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. In: Working Notes 6th Workshop of the Cross-Language Evaluation Forum, CLEF, Vienna (2005)
8. Hackl, R., Mandl, T., Womser-Hacker, C.: Mono- and Cross-lingual Retrieval Experiments at the University of Hildesheim. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 165–169. Springer, Heidelberg (2005)
9. Larson, R.: Chesire II at GeoCLEF: Fusion and Query Expansion for GIR. In: Working Notes 6th Workshop of the Cross-Language Evaluation Forum, Vienna (2005)
10. Mandl, T., Schneider, R., Strötgen, R.: A Fast Forward Approach to Cross-lingual Question Answering for English and German. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 332–336. Springer, Heidelberg (2006)

⁵ <http://wortschatz.uni-leipzig.de/>

A WordNet-Based Indexing Technique for Geographical Information Retrieval

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis

Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia, Spain
{dbuscaldi,proso,esanchis}@dsic.upv.es

Abstract. This paper presents an indexing technique based on WordNet synonyms and holonyms. This technique has been developed for the Geographical Information Retrieval task. It may help in finding implicit geographic information contained in texts, particularly if the indication of the containing geographical entity is omitted. Our experiments were carried out with the Lucene search engine over the GeoCLEF 2006 set of topics. Results show that expansion can improve recall in some cases, although a specific ranking function is needed in order to obtain better results in terms of precision.

1 Introduction

Nowadays, many documents in the web or in digital libraries contain some kind of geographical information. News stories often contain a reference that indicates the place where an event took place. Nevertheless, the correct identification of the locations to which a document refers to is not a trivial task. Explicit information about areas including the cited geographical entities is usually missing from texts (e.g. usually *France* is not named in a news related to *Paris*). Moreover, using text strings in order to identify a geographical entity creates problems related to ambiguity, synonymy and names changing over time.

Ambiguity and synonymy are well-known problems in the field of Information Retrieval. The use of semantic knowledge may help to solve these problems, even if no strong experimental results are yet available in support of this hypothesis. Some results [1] show improvements by the use of semantic knowledge; others do not [7]. The most common approaches make use of standard keyword-based techniques, improved through the use of additional mechanisms such as document structure analysis and automatic query expansion.

We investigated the use of automatic query expansion by means of WordNet [6] meronyms and synonyms in our previous participation to the GeoCLEF, but the proposed technique did not obtain good results [5,3]. Although there are some effective query expansion techniques [4] that can be applied to the geographical domain, we think that the expansion of the queries with synonyms and meronyms does not fit with the characteristics of the GeoCLEF task. Other methods using thesauri with synonyms for general domain IR also did not achieve promising results [8].

In our work for GeoCLEF 2006 we focused on the use of WordNet in the indexing phase, specifically for the expansion of index terms by means of synonyms and holonyms. We used the subset of the WordNet ontology related to the geographical domain.

2 WordNet-Based Index Terms Expansion

The expansion of index terms is a method that exploits the *holonymy* relationship of the WordNet ontology. A concept *A* is *holonym* of another concept *B* if *A* contains *B*, or viceversa *B* is part of *A* (*B* is also said to be *meronym* of *A*). Therefore, our idea is to add to the geographical index terms the informations about their holonyms, such that a user looking information about *Spain* will find documents containing *Valencia*, *Madrid* or *Barcelona* even if the document itself does not contain any reference to Spain.

We used the well-known Lucene¹ search engine. Two indices are generated for each text during the indexing phase: a *geo* index, containing all the geographical terms included in the text and also those obtained through WordNet, and a *text* index, containing the stems of text words that are not related to geographical entities. Thanks to the separation of the indices, a document containing “John Houston” will not be retrieved if the query contains “Houston”, the city in Texas. The adopted weighting scheme is the usual $tf \cdot idf$. The geographical terms in the text are identified by means of a Named Entity (NE) recognizer based on maximum entropy², and put into the *geo* index, together with all its synonyms and holonyms obtained from WordNet.

For instance, consider the following text:

“A federal judge in Detroit struck down the National Security Agency’s domestic surveillance program yesterday, calling it unconstitutional and an illegal abuse of presidential power.”

The NE recognizer identifies *Detroit* as a geographical entity. A search for Detroit synonyms in WordNet returns {*Detroit*, *Motor city*, *Motown*}, while its holonyms are:

```
-> Michigan, Wolverine State, Great Lakes State, MI
    -> Midwest, middle west, midwestern United States
        -> United States, United States of America, U.S.A., USA,
            U.S., America
            -> North America
                -> northern hemisphere
                -> western hemisphere, occident, New World
                -> America
```

¹ <http://lucene.jakarta.org>

² Freely available from the OpenNLP project: <http://opennlp.sourceforge.net>

Therefore, the following index terms are put into the *geo* index: { *Michigan*, *Wolverine State*, *Great Lakes State*, *MI*, *Midwest*, *middle west*, *midwestern United States*, *United States*, *United States of America*, *U.S.A.*, *USA*, *U.S.*, *America*, *North America*, *northern hemisphere*, *western hemisphere*, *occident*, *New World* }.

3 Experimental Results

The runs submitted at GeoCLEF 2006 were four, two with the WordNet-based system and the other ones with the “clean” system (i.e. without the expansion of index terms). The runs were the mandatory “title-description” and “title-description-narrative” for each of the two systems. For every query the top 1000 ranked documents have been returned.

In table 1 we show the recall and average precision values obtained. Recall has been calculated for each run as the number of relevant documents retrieved divided by the number of relevant documents in the collection (378). The average precision is the non-interpolated average precision calculated for all relevant documents, averaged over queries.

The results obtained in term of precision show that non-WordNet runs are better than the other ones, particularly for the all-fields run *rfiaUPV02*. However, as we expected, we obtained an improvement in recall for the WordNet-based system, although the improvement was not so significant as we hoped (about 1%).

Table 1. Average precision and recall values obtained for the four runs. WN: tells whether the run uses WordNet or not.

run	WN	avg. precision	recall
rfiaUPV01	no	25.07%	78.83%
rfiaUPV02	no	27.35%	80.15%
rfiaUPV03	yes	23.35%	79.89%
rfiaUPV04	yes	26.60%	81.21%

In order to better understand the obtained results, we analyzed the topics in which the two systems differ more (in terms of recall). Topics 40 and 48 resulted the worst ones for the WordNet based system. The explication is that topic 40 does not contain any name of geographical place (“*Cities near active volcanoes*”); topic 48 contains references to places (*Greenland* and *Newfoundland*) for which WordNet provides little information.

On the other hand, the system based on index term expansion performed particularly well for topics 27, 37 and 44. These topics contain references to countries and regions (*Western Germany* for topic 27, *Middle East* in the case of 37 and *Yugoslavia* for 44) for which WordNet provides a rich information in terms of meronyms.

4 Conclusions and Further Work

The obtained results show that the expansion of index terms by means of WordNet holonyms can improve slightly the recall. However, a better ranking function needs to be implemented in order to obtain also an improvement in precision. Further investigation directions will include the implementation of the same method with a richer (in terms of coverage of geographical places) resource, an ontology we are currently developing using the GNS and GNIS gazetteers together with WordNet itself and Wikipedia [2], as the experimentation of various ranking functions that weight differently the geographical terms with respect to the non-geographical ones.

Acknowledgments

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work. This paper is a revised version of the work titled “WordNet-based Index Terms Expansion for Geographical Information Retrieval” included in the CLEF 2006 Working Notes.

References

1. Bo-Yeong, K., Hae-Jung, K., Sang-Lo, L.: Performance analysis of semantic indexing in text retrieval. In: Gelbukh, A. (ed.) *CICLing 2004*. LNCS, vol. 2945, Springer, Heidelberg (2004)
2. Buscaldi, D., Rosso, P., Peris, P.: Inferring geographical ontologies from multiple resources for geographical information retrieval. In: *Proceedings of the 3rd GIR Workshop, SIGIR 2006*, Seattle, WA (2006)
3. Buscaldi, D., Rosso, P., Sanchis, E.: Using the wordnet ontology in the geoclef geographical information retrieval task. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-based spatial query expansion in information retrieval. In: *Proceedings of the ODBASE 2005 conference* (2005)
5. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P.: Geoclef: the clef 2005 cross-language geographic information retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005*. LNCS, vol. 4022, Springer, Heidelberg (2006)
6. Miller, G.A.: Wordnet: A lexical database for English (chapter 11). *Communications of the ACM* 38, 39–41 (1995)
7. Rosso, P., Ferretti, E., Jiménez, D., Vidal, V.: Text categorization and information retrieval using wordnet senses. In: Gelbukh, A. (ed.) *CICLing 2004*. LNCS, vol. 2945, Springer, Heidelberg (2004)
8. Voorhees, E.: Query expansion using lexical-semantic relations. In: *Proceedings of the ACM SIGIR 1994*, ACM Press, New York (1994)

University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval

Claudia Hauff, Dolf Trieschnigg, and Henning Rode

University of Twente
Human Media Interaction Group
Enschede, The Netherlands
`{hauffc,trieschn,rodeh}@ewi.utwente.nl`

Abstract. This paper describes the approach of the University of Twente at its first participation in GeoCLEF. A large effort went into the construction of a geographic thesaurus which was utilized to add geographic knowledge to the documents and queries. Geographic filtering was applied to the results returned from a retrieval by content run. Employing such a geographic knowledge base however showed no added value - the content-only baseline outperformed all geographically filtered runs.

1 Introduction

During GeoCLEF's pilot track in 2005 became clear that incorporating geographical knowledge into an information retrieval (IR) system is not always beneficial. The best performance of GeoCLEF 2005 was achieved using standard keyword search techniques [1,2,3]. Only Metacarta [4]'s approach using geographic bounding boxes did outperform their keyword-only approach. However, the resulting mean average precision did not exceed 17% and was far from the best submissions (36% and 39% respectively).

Despite the disappointing results of those efforts to incorporate some spatial awareness in IR systems, we believe that adding knowledge about locality can improve search performance. In our CLEF submission of this year, we have confined ourselves to the monolingual task and have only worked with the English queries and documents.

Our approach can be summarized as follows:

1. Carry out document retrieval to find “topically relevant” documents. For example, for the topic “Car bombings near Madrid” this step should result in a ranked list of documents discussing “car bombings”, not necessarily near Madrid.
2. Filter this ranked list based on “geographical relevance”. For each topically relevant document, determine if it is also geographically relevant. If not, it is removed from the list.

The outline of this paper is as follows. In Section 2 the geographic gazetteer (a thesaurus like resource) we created is discussed. The following section describes the process of tagging the document collection and the queries (Section 3) with geographical information. Section 4 describes the document retrieval and filtering processes. Finally, in Section 5 the experiment results are listed followed by a discussion and conclusion in Section 6.

2 The Gazetteer

The gazetteer we used lists geographical references (strings of text) and links them to geographical locations, defined through longitude and latitude values. It also provides information about parent-child relationships between those references. A parent is defined as a region or a country, hence information such as “Madrid lies in Spain which is part of Europe” can be found in the gazetteer.

Our gazetteer was built up from freely available resources. To achieve world coverage, the following gazetteers were combined:

- GEOnet Names Server (<http://earth-info.nga.mil/gns/html/>),
- the Geographics Names Information System (<http://geonames.usgs.gov/stategaz/>),
- the World Gazetteer (<http://www.world-gazetteer.com/>).

An inference mechanism was employed to maximize the amount of parent-child information. Missing parent information of an entry was inferred from agreeing parent information of nearby location entries. The coverage of the merged gazetteer though is not uniform: whereas the USA and Western Europe are well represented, other regions - such as Canada, Northern Africa and a large part of Asia - are barely covered. Figure 1 shows the coverage of the gazetteer. Grid regions with few gazetteer entries are green (light), while red (darker) areas are densely covered.

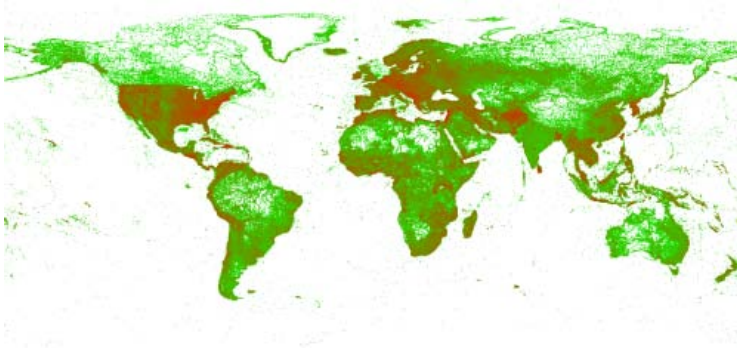


Fig. 1. Location density of the merged gazetteer

3 Corpus and Query Processing

In the preprocessing stage, the geographical range of each document is determined in a two-step process. First, potential location phrases in documents are identified by searching for the longest phrases of capitalized letter strings in each sentence. One additional rule is applied: if two found phrases are only separated by ‘of’, these phrases are treated as one (for example “Statue of Liberty”). In the second step, the list of potential locations found is matched against the geographic references in the gazetteer.

Contrary to last year, the GeoCLEF topics were not provided with geographic tags. We processed the topics’ title section manually and tagged the location names and the type of spatial relation (around, north, east, etcetera). Given a query, its potential locations were extracted. For a tagged query this is merely the text between the location tags. For an untagged query all capitalized letter phrases were treated as location candidates and matched against the gazetteer. If the gazetteer did not contain any candidate, the extracted phrases were used as Wikipedia queries and the returned wiki page was geotagged the same way as the corpus documents. Additionally, if the found location was a country, its boundaries (minimum and maximum latitude/longitude pairs in the gazetteer) were applied as location coordinate restrictions.

4 Document Retrieval and Filtering

The document collection was indexed with the Lemur Toolkit for Language Modeling and for retrieval purposes the language modelling approach with Jelineck-Mercer smoothing was chosen. Given a query the ranked list of results of a retrieval by content run is returned and subsequently filtered to remove documents outside the desired geographical scope. The locations found in the document are matched against the coordinate restrictions obtained from the query. A document is removed from the result list, if it does not contain any locations which fulfil the query coordinate restrictions.

For queries without coordinate restrictions, the sets of query and document locations are split into parents sets Q_p (query parents) and D_p (document parents). The children sets Q_c (query children) and D_c (document children) are the location names that appear in the gazetteer but not as a parent. In order to determine geographical relevance the intersection sets $I_p = Q_p \cap D_p$ and $I_c = Q_c \cap D_c$ are evaluated. If $Q_x \neq \emptyset$ with $x = \{p, c\}$, then $I_x \neq \emptyset$ must hold in order for the document to be geographically relevant.

5 Experiments and Results

We tested different variations of the usage of title, description and narrative as well as merging the filtered results with the content-only ranking by adding the top filtered-out results at the end of the ranking. The results are given in Table 1. The baseline run in each case is the content-only run.

Table 1. Results (Mean Average Precision) for the English task of GeoCLEF

run id	title	desc.	narr.	geo	merged	map
baseline	x					17.45%
utGeoTIB	x			x		16.23%
utGeoTIBm	x			x	x	17.18%
baseline	x	x				15.24%
utGeoTdIB	x	x		x		7.32%
baseline	x	x	x			18.75%
utGeoTdnIB	x	x	x	x		11.34%
utGeoTdnIBm	x	x	x	x	x	16.77%

6 Discussion and Conclusion

The results show no improvement over the baseline by the addition of geographical knowledge; on the contrary, the performance degrades significantly when including the description or narrative of a topic in a query. A manual evaluation of the relevant documents of the first eight GeoCLEF 2006 topics revealed, that the exact location phrases (e.g. *in the northern part of Iraq*) mentioned in the title query also occur in almost all relevant documents. This makes a geographically enhanced approach unnecessary and also explains the similar results between the baseline and the geographically filtered results for the title queries.

The performance drop of the description and narrative queries is suspected to be due to the fact that many queries contain a long list of possible location terms within them. For retrieval purposes, the location terms within the queries are treated like every other query keyword. However, they are different in the sense that their term frequency within the documents is of no importance; mentioning the location term once within the document already determines the location.

In conclusion, we are still unable to provide conclusive evidence for or against the usage of a geographical knowledge base in the ad hoc information retrieval task. In future work we will evaluate the quality of our geotagging process and its influence on retrieval performance. Furthermore probabilistic geotagging and retrieval models will be investigated.

References

1. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: Geoclef: the clef 2005 cross-language geographic information retrieval track overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
2. Gey, F., Petras, V.: Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
3. Guillén, R.: CSUSM Experiments in GeoCLEF2005: Monolingual and Bilingual Tasks. In: Working Notes for the CLEF 2005 Workshop (2005)
4. Kornai, A.: MetaCarta at GeoCLEF 2005. In: GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview (2005)

TALP at GeoCLEF 2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus

Daniel Ferrés and Horacio Rodríguez

TALP Research Center, Software Department
Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08043 Barcelona, Spain
{dferres, horacio}@lsi.upc.edu
<http://www.lsi.upc.edu/~nlp>

Abstract. This paper describes our experiments in Geographical Information Retrieval (GIR) in the context of our participation in the CLEF 2006 GeoCLEF Monolingual English task. Our system, named TALP-GeoIR, follows a similar architecture of the GeoTALP-IR system presented at GeoCLEF 2005 with some changes in the retrieval modes and the Geographical Knowledge Base (KB).

The system has four phases performed sequentially: i) a Keyword Selection algorithm based on a linguistic and geographical analysis of the topics, ii) a geographical retrieval with *Lucene*, iii) a document retrieval task with the *JIRS* Passage Retrieval (PR) software, and iv) a Document Ranking phase. A Geographical KB has been built using a set of publicly available geographical gazetteers and the *Alexandria Digital Library (ADL) Feature Type Thesaurus*.

In our experiments we have used *JIRS*, a state-of-the-art PR system for Question Answering, for the GIR task. We also have experimented with an approach using both *JIRS* and *Lucene*. In this approach *JIRS* was used only for textual document retrieval and *Lucene* was used to detect the geographically relevant documents. These experiments show that applying only *JIRS* we obtain better results than combining *JIRS* and *Lucene*.

1 Introduction

This paper describes our experiments in Geographical Information Retrieval (GIR) in the context of our participation in the CLEF 2006 GeoCLEF Monolingual English task.

GeoCLEF is a cross-language geographical retrieval evaluation task at the CLEF 2006 campaign [4]. Like the first GeoCLEF GIR task at CLEF 2005 [5], the goal of the GeoCLEF task is to find as many relevant documents as possible from the document collections using a topic set with spatial user needs.

Our GIR system is a modified version of the system presented in GeoCLEF 2005 [2] with some changes in the retrieval modes and the Geographical

Knowledge Base (KB). The system has four phases performed sequentially: i) a Keywords Selection algorithm based on a linguistic and geographical analysis of the topics, ii) a geographical document retrieval with *Lucene*, iii) a textual document retrieval with the *JIRS* Passage Retrieval (PR) software, and iv) a Document Ranking phase. A Geographical KB has been build using a set of publicly available geographical gazetteers and the *Alexandria Digital Library (ADL) Feature Type Thesaurus*.

In this paper we present the overall architecture of the TALP-GeoIR system and we describe briefly its main components. We also present our experiments, results, and conclusions in the context of the CLEF 2006 GeoCLEF Monolingual English task evaluation.

2 System Description

The system architecture has four phases that are performed sequentially: Topic Analysis, Geographical Retrieval, Textual Retrieval, and Document Ranking. Previously, a Collection Processing phase has been applied over the textual collections.

2.1 Collection Processing

We processed the entire English collections: Glasgow Herald 1995 and Los Angeles Times 1994 with linguistic tools (described in the next sub-section) to mark the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process, the collection is analyzed with a Geographical KB (described in the next sub-section). This information was used to build two indexes: one with the geographical information of the documents and another one with the textual and geographical information of the documents. We have used two Information Retrieval (IR) systems to create these indexes: *Lucene*¹ for the geographical index and *JIRS*² for the textual and geographical index (see a sample of both indexes in Figure 1). These indexes are described below:

- **Geographical Index:** this index contains the geographical information of the documents and its Named Entities. The Geographical index contains the following fields for each document:
 - **docid:** This field stores the document identifier.
 - **ftt:** This field indexes the feature type of each geographical name and the Named Entity classes of all the NEs appearing in the document.
 - **geo:** This field indexes the geographical names and the Named Entities of the document. It also stores the geographical information (feature type, hierachical ancestors' path, and coordinates) about the place names. Even if the place is ambiguous all the possible referents are indexed.

¹ **Lucene.** <http://jakarta.apache.org/lucene>

² **JIRS.** <http://jirs.dsic.upv.es>

- **Textual and Geographical Index:** This index stores the lemmatized content of the document and the geographical information (feature type, hierarchical ancestors’ path, and coordinates) about the geographic place names appearing in the text. If the geographical place is ambiguous then this information is not added to the indexed content.

System	Indexed Content	
Lucene	docid	GH950102000000
	ftt	regions@land_regions@continents administrative_areas@political_areas@countries_1st_order_divisions administrative_areas@populated_places@cities administrative_areas@political_areas@countries ...
	geo	Europe Asia@Western_Asia@Saudi_Arabia@Hejaz@24.5_38.5 America@Northern_America@United_States@South_Carolina@Lodge@32.9817_-80.952 America@Northern_America@United_States@38.91_-96.19 ...
JIRS	... the role of the wheel in lamatrekking , and where be the good place to air your string vest. pity the crew who accompany him on his travel as sayle of Arabia <i>countries_1st_order_divisions</i> Asia Western_Asia Kuwait Arabia 25.0_45.0 along the Hejaz <i>countries_1st_order_divisions</i> Asia Western_Asia Saudi_Arabia Hejaz 24.5_38.5 railway line from Aleppo <i>countries_1st_order_divisions</i> Asia Middle_East Syria Aleppo 36.0_37.0 in Northern_Syria <i>countries</i> Asia Middle_East Syria 35.0_38.0 to Aqaba <i>cities</i> Asia Western_Asia Jordan Maán Aqaba 29.517_35 in Jordan <i>countries</i> Asia Western_Asia Jordan 31.0_36.0 . as he journey through the searing heat in an age East German...	

Fig. 1. Samples of an indexed document with *Lucene* and *JIRS*

2.2 Topic Analysis

The goal of this phase is to extract all the relevant keywords (along with linguistic and geographic analyses) from the topics. These keywords are then used by the document retrieval phases. The Topic Analysis phase has three main components: a Linguistic Analysis, a Geographical Analysis, and a Keyword Selection algorithm.

Linguistic Analysis. This process extracts lexico-semantic and syntactic information using these Natural Language Processing tools: i) *TnT* part-of-speech tagger [1], ii) *WordNet 2.0 lemmatizer*, iii) *Spear*³ (a modified version of the *Collins parser* [3]), and iv) a *Maximum Entropy* based Named Entity Recognizer and Classifier (NERC) trained with the CONLL-2003 shared task English data set [9].

³ *Spear*. <http://www.lsi.upc.edu/~surdeanu/spear.html>

Geographical Analysis. The Geographical Analysis is applied to the NEs from the title, description, and narrative tags that have been classified as *location* or *organization* by the NERC tool. This analysis has two components:

- **Geographical Knowledge Base:** this component has been built joining four geographical gazetteers:
 1. *GEOnet Names Server (GNS)*⁴: a gazetteer covering worldwide excluding the United States and Antarctica, with 5.3 million entries.
 2. *Geographic Names Information System (GNIS)*⁵: we used the *GNIS Concise Features*⁶ data set with 39,906 entries of United States.
 3. *GeoWorldMap World Gazetteer*⁷: a gazetteer with approximately 40,594 entries of the most important countries, regions and cities of the world.
 4. *World Gazetteer*⁸: a gazetteer with approximately 171,021 entries of towns, administrative divisions and agglomerations with their features and current population. From this gazetteer we added only the 29,924 cities with more than 5,000 inhabitants.
- **Geographical Feature Type Thesaurus:** the feature type thesaurus of our Geographical KB is the *ADL Feature Type Thesaurus (ADLFTT)*. The *ADL Feature Type Thesaurus* is a hierarchical set of geographical terms used to type named geographic places in English [6]. Both *GNIS* and *GNS* gazetteers have been mapped to the *ADLFTT*, with a resulting set of 575 geographical types. Our *GNIS* mapping is similar to the one exposed in [6].

Topic Keywords Selection. This algorithm extracts the most relevant keywords of each topic (see an example in Figure 2). The algorithm was designed for GeoCLEF 2005 [2]. The algorithm is applied after a linguistic and geographical analysis and has the following steps:

1. All the punctuation symbols and stop-words are removed from the analysis of the title, description and narrative tags.
2. All the words from the title tag are obtained.
3. All the Noun Phrase base chunks from the description and narrative tags that contain a word with a lemma that appears in one or more words from the title are extracted
4. The words that pertain to the chunks extracted in the previous step and do not have a lemma appearing in the words of the title are extracted.

Once the keywords are extracted, three different Keyword Sets (KS) are created:

- **All:** All the keywords extracted from the topic tags.
- **Geo:** Geographical places or feature types appearing in the topic tags.
- **NotGeo:** All the keywords extracted from the topic tags that are not geographical place names or geographical types.

⁴ **GNS.** <http://earth-info.nga.mil/gns/html/namefiles.htm>

⁵ **GNIS.** http://geonames.usgs.gov/domestic/download_data.htm

⁶ Contains large features that should be labeled on maps with a scale of 1:250,000.

⁷ **GeoWorldMap** (Geobytes Inc.). <http://www.geobytes.com>

⁸ **World Gazetteer.** <http://www.world-gazetteer.com>

Topic	EN-title	Wine regions around rivers in Europe
	EN-desc	Documents about wine regions along the banks of European rivers.
	EN-narr	Relevant documents describe a wine region along a major river in European countries. To be relevant the document must name the region and the river.
Keywords Set (KS)	Not Geo	wine European
	Geo	Europe#location#regions@land_regions@continents#Europe regions hydrographic_features@streams@rivers
	All	wine regions rivers European Europe

Fig. 2. Keyword sets sample of Topic 026

2.3 Geographical Document Retrieval with *Lucene*

Lucene is used to retrieve geographically relevant documents given a specific Geographical IR query. *Lucene* uses the standard tf-idf weighting scheme with the cosine similarity measure and allows ranked and boolean queries. We used boolean queries with a *Relaxed geographical search policy* (see [2] for more details). This search policy allows the system to retrieve all the documents that have a token that matches totally or partially (a sub-path) the geographical keyword. As an example, the keyword `America@Northern_America@United_States` will retrieve all the U.S. places (e.g. `America@Northern_America@United_States@Ohio`).

2.4 Document Retrieval Using the *JIRS* Passage Retriever

The *JIRS* Passage Retrieval System [8] is used to retrieve relevant documents related to a GIR query. *JIRS* is a Passage Retriever specially designed for Question Answering (QA). This system gets passages with a high similarity between the largests n-grams of the question and the ones in the passage.

We used *JIRS* considering a topic keyword set as a question. Then, we retrieved passages using the n-gram distance model of *JIRS* with a length of 11 sentences per passage. We obtained the first 100.000 top-scored passages per topic. Finally, a process selects the relevant documents from the set of retrieved passages. Two document scoring strategies were used:

- **Best:** The document score is the score of the top-scored passage in the set of the retrieved passages that belong to this document.
- **Accumulative:** The document score is the sum of the scores of all the retrieved passages that belong to this document.

2.5 Document Ranking

This component ranks the documents retrieved by *Lucene* and *JIRS*. First, the top-scored documents retrieved by *JIRS* that appear in the document set retrieved by

Lucene are selected. Then, if the set of selected documents is less than 1,000 the top-scored documents of *JIRS* that not appear in the document set of *Lucene* are selected with a lower priority than the previous ones. Finally, the first 1,000 top-scored documents are selected. On the other hand, when the system uses only *JIRS* for retrieval only the first 1,000 top-scored documents by *JIRS* are selected.

3 Experiments

We designed a set of five experiments that consist in applying different IR systems, query keyword sets, and tags to an automatic GIR system (see Table 1). Basically, these experiments can be divided in two groups depending on the retrieval engines used:

- **JIRS.** Two baseline experiments have been done in this group: the runs *TALP-GeoIRTD1* and *TALPGeoIRTDN1*. These runs differ uniquely in the use of the narrative tag in the second one. Both runs use one retrieval system, *JIRS*, and they use all the keywords to perform the query. The experiment *TALP-GeoIRTDN3* is similar to the previous ones but uses a Cumulative scoring strategy to select the documents with *JIRS*.
- **JIRS & Lucene.** The runs *TALPGeoIRTD2* and *TALPGeoIRTDN2* use *JIRS* for textual document retrieval and *Lucene* for geographical document retrieval. Both runs use the *Geo* keywords set for *Lucene* and the *NotGeo* keywords set for *JIRS*.

Table 1. Description of the experiments at GeoCLEF 2006

Automatic Runs	Tags	IR System	JIRS KS	Lucene KS	<i>JIRS</i> Score
TALPGeoIRTD1	TD	JIRS	All	-	Best
TALPGeoIRTD2	TD	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN1	TDN	JIRS	All	-	Best
TALPGeoIRTDN2	TDN	JIRS+Lucene	NotGeo	Geo	Best
TALPGeoIRTDN3	TDN	JIRS	All	-	Cumulative

4 Results

The results of the TALP-GeoIR system at the CLEF 2006 GeoCLEF Monolingual English task are summarized in Table 2. This table has the following IR measures for each run: *Average Precision*, *R-Precision*, and *Recall*.

The results show a substantial difference between the two sets of experiments. The runs that use only *JIRS* have a better *Average Precision*, *R-Precision*, and *Recall* than the ones that use *JIRS* and *Lucene*. The run with the best *Average Precision* is *TALPGeoIRTD1* with 0.1342. The best *Recall* measure is obtained by

the run *TALPGeoIRTDN1* with a 68.78% of the relevant documents retrieved. This run has the same configuration that the *TALPGeoIRTD1* run but uses the narrative tag. Finally, we obtained poor results in comparison with the mean average precision (0.1975) obtained by all the systems that participated in the GeoCLEF 2006 Monolingual English task.

Table 2. TALP-GeoIR results at GeoCLEF 2006 Monolingual English task

Automatic Runs	Tags	IR System	AvgP.	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	JIRS	0.1342	0.1370	60.84%	230/378
TALPGeoIRTD2	TD	JIRS+Lucene	0.0766	0.0884	32.53%	123/378
TALPGeoIRTDN1	TDN	JIRS	0.1179	0.1316	68.78%	260/378
TALPGeoIRTDN2	TDN	JIRS+Lucene	0.0638	0.0813	47.88%	181/378
TALPGeoIRTDN3	TDN	JIRS	0.0997	0.0985	64.28%	243/378

5 Conclusions

We have applied *JIRS*, a state-of-the-art PR system for QA, to the GeoCLEF 2006 Monolingual English task. We also have experimented with an approach using both *JIRS* and *Lucene*. In this approach *JIRS* was used only for textual document retrieval and *Lucene* was used to detect the geographically relevant documents. The approach with only *JIRS* was better than the one with *JIRS* and *Lucene* combined.

Comparatively with the Mean Average Precision (MAP) of all the runs participating at GeoCLEF 2006 Monolingual English task our MAP is low. This fact can be due to several reasons: i) the *JIRSPR* system may be was not used appropriately or is not suitable for GIR, ii) our system is not dealing with geographical ambiguities, iii) the lack of textual query expansion methods, iv) the need of Relevance Feedback methods, and v) errors in the Topic Analysis phase.

As a future work we propose the following improvements to the system: i) the resolution of geographical ambiguity problems applying toponym resolution algorithms, ii) to apply some query expansion methods, iii) to apply Blind Feedback techniques.

Acknowledgments

This work has been partially supported by the European Commission (CHIL, IST-2004-506909). Daniel Ferrés is supported by a UPC-Recerca grant from Universitat Politècnica de Catalunya (UPC). TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

References

1. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the 6th Applied NLP Conference (ANLP-2000), Seattle, WA, United States (2000)
2. Ferrés, D., Ageno, A., Rodríguez, H.: The GeoTALP-IR System at GeoCLEF-2005: Experiments Using a QA-based IR System, Linguistic Analysis, and a Geographical Thesaurus. In: Peters, et al. (eds.) [7] (2005)
3. Ferrés, D., Kanaan, S., González, E., Ageno, A., Rodríguez, H., Surdeanu, M., Turmo, J.: TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints. In: Proceedings of the Text Retrieval Conference (TREC-2004) (2005)
4. Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Nunzio, G.M.D., Ferro, N.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Proceedings of the Cross Language Evaluation Forum 2006. LNCS, Springer, Heidelberg (2007) (in this volume)
5. Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) [7], CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
6. Hill, L.L.: Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 280–290. Springer, Heidelberg (2000)
7. Peters, C., Gey, F.C., Gonzalo, J., Jones, G.J.F., Müller, H., Kluck, M., Magnini, B., Müller, H., de Rijke, M.: Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., Müller, H., de Rijke, M. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
8. Gómez Soriano, J.M., Montes-y-Gómez, M., Arnal, E.S., Rosso, P.: A Passage Retrieval System for Multilingual Question Answering. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 443–450. Springer, Heidelberg (2005)
9. Sang, E.F.T.K., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Daelemans, W., Osborne, M. (eds.) Proceedings of CoNLL-2003, pp. 142–147. Edmonton, Canada (2003)

GeoCLEF Text Retrieval and Manual Expansion Approaches

Ray R. Larson and Fredric C. Gey

School of Information
University of California, Berkeley, USA
`ray@sims.berkeley.edu`, `gey@berkeley.edu`

Abstract. In this paper we will describe the Berkeley approaches to the GeoCLEF tasks for CLEF 2006. This year we used two separate systems for different tasks. Although of the systems both use versions of the same primary retrieval algorithm they differ in the supporting text pre-processing tools used.

1 Introduction

This paper describes the retrieval algorithms and evaluation results for Berkeley's official submissions for the GeoCLEF track. Two separate systems were used for our runs, although both used the same basic algorithm for retrieval. Instead of the automatic expansion used in last year's GeoCLEF, this year we used manual expansion for a selected subset of queries for only 2 out of the 18 runs submitted. The remainder of the runs were automatic without manual intervention in the queries (or translations). We submitted 12 Monolingual runs (2 German, 4 English, 2 Spanish, and 4 Portuguese) and 6 Bilingual runs (2 English $\Rightarrow$ German, 2 English $\Rightarrow$ Spanish, and 2 English $\Rightarrow$ Portuguese). We did not submit any Biligual X $\Rightarrow$ English runs.

This paper first describes the retrieval algorithms used for our submissions, followed by a discussion of the processing used for the runs. We then examine the results obtained for our official runs, and finally present conclusions and future directions for GeoCLEF participation.

2 The Retrieval Algorithms

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions was originally developed by Cooper, et al. [5]. The LR model of probabilistic IR estimates the probability of relevance for each document based on statistics about a document collection and a set of queries in combination with a set of weighting coefficients for those statistics. The statistics to be used and the values of the coefficients are obtained from regression analysis of a sample of a collection (or similar test collection) for some set of queries where relevance and non-relevance has been determined. More formally, given

a particular query and a particular document in a collection $P(R | Q, D)$ is calculated and the documents or components are presented to the user ranked in order of decreasing values of that probability. To avoid invalid probability values, the usual calculation of $P(R | Q, D)$ uses the “log odds” of relevance given a set of S statistics, s_i , derived from the query and database, such that:

$$\log O(R | Q, D) = b_0 + \sum_{i=1}^S b_i s_i \quad (1)$$

where b_0 is the intercept term and the b_i are the coefficients obtained from the regression analysis of the sample collection and relevance judgements. The final ranking is determined by the conversion of the log odds form to probabilities:

$$P(R | Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}} \quad (2)$$

2.1 TREC2 Logistic Regression Algorithm

For GeoCLEF we used a version the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a number of years[3]. We used two different implementations of the algorithm. One was in stand-alone experimental software developed by Aitao Chen, and the other in the Cheshire II information retrieval system. Although the basic behaviour of the algorithm is the same for both systems, there are differences in the sets of pre-processing and indexing elements used in retrieval. One of the primary differences is the lack of compounding for German documents and query terms in the Cheshire II system. The formal definition of the TREC2 Logistic Regression algorithm used is:

$$\begin{aligned} \log O(R|C, Q) &= \log \frac{p(R|C, Q)}{1 - p(R|C, Q)} = \log \frac{p(R|C, Q)}{p(\bar{R}|C, Q)} \\ &= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \frac{qt f_i}{ql + 35} \\ &\quad + c_2 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \log \frac{t f_i}{cl + 80} \\ &\quad - c_3 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \log \frac{ct f_i}{N_t} \\ &\quad + c_4 * |Q_c| \end{aligned} \quad (3)$$

where C denotes a document component (i.e., an indexed part of a document which may be the entire document) and Q a query, R is a relevance variable,

$p(R|C, Q)$ is the probability that document component C is relevant to query Q ,

$p(\overline{R}|C, Q)$ the probability that document component C is *not relevant* to query Q , which is $1.0 - p(R|C, Q)$
 $|Q_c|$ is the number of matching terms between a document component and a query,
 qtf_i is the within-query frequency of the i th matching term,
 tf_i is the within-document frequency of the i th matching term,
 ctf_i is the occurrence frequency in a collection of the i th matching term,
 ql is query length (i.e., number of terms in a query like $|Q|$ for non-feedback situations),
 cl is component length (i.e., number of terms in a component), and
 N_t is collection length (i.e., number of terms in a test collection).
 c_k are the k coefficients obtained through the regression analysis.

If stopwords are removed from indexing, then ql , cl , and N_t are the query length, document length, and collection length, respectively. If the query terms are re-weighted (in feedback, for example), then qtf_i is no longer the original term frequency, but the new weight, and ql is the sum of the new weight values for the query terms. Note that, unlike the document and collection lengths, query length is the “optimized” relative frequency without first taking the log over the matching terms.

The coefficients were determined by fitting the logistic regression model specified in $\log O(R|C, Q)$ to TREC training data using a statistical software package. The coefficients, c_k , used for our official runs are the same as those described by Chen[1]. These were: $c_0 = -3.51$, $c_1 = 37.4$, $c_2 = 0.330$, $c_3 = 0.1937$ and $c_4 = 0.0929$. Further details on the TREC2 version of the Logistic Regression algorithm may be found in Cooper et al. [4].

2.2 Blind Relevance Feedback

In addition to the direct retrieval of documents using the TREC2 logistic regression algorithm described above, we have implemented a form of “blind relevance feedback” as a supplement to the basic algorithm. The algorithm used for blind feedback was originally developed and described by Chen [2]. Blind relevance feedback has become established in the information retrieval community due to its consistent improvement of initial search results as seen in TREC, CLEF and other retrieval evaluations [6]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [8].

Blind relevance feedback is typically performed in two stages. First, an initial search using the original topic statement is performed, after which a number of terms are selected from some number of the top-ranked documents (which are presumed to be relevant). The selected terms are then weighted and then merged with the initial query to formulate a new query. Finally the reweighted and expanded query is submitted against the same collection to produce a final ranked list of documents. For GeoCLEF run using the Cheshire system this year, we chose to use the top 13 terms from 16 top-ranked documents, based on analysis of our 2005 GeoCLEF results.

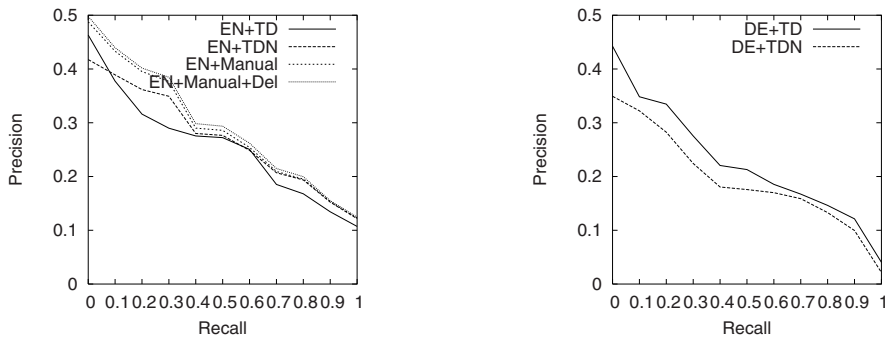


Fig. 1. Berkeley Monolingual Runs – English (left) and German (right)

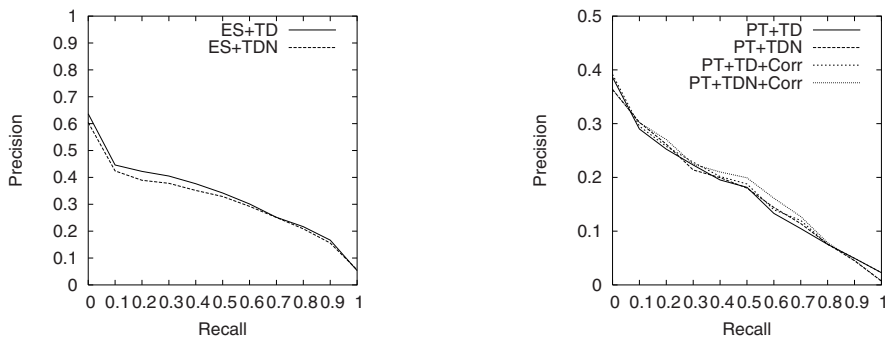


Fig. 2. Berkeley Monolingual Runs – Spanish (left) and Portuguese (right)

3 Approaches for GeoCLEF

In this section we describe the specific approaches taken for our submitted runs for the GeoCLEF task. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

3.1 Indexing and Term Extraction

The standalone version treats all text as a single “bag of words” that is extracted and indexed. For German documents it uses a custom “decompounding” algorithm to extract component terms from German compounds.

The Cheshire II system uses the XML structure and extracts selected portions of the record for indexing and retrieval. However, for our official runs this year we used the full text, and avoided merging results from multiple elements.

Because there was no explicit tagging of location-related terms in the collections used for GeoCLEF, we applied the above approach to the “TEXT”, “LD”, and “TX” elements of the records of the various collections. The part of news articles normally called the “dateline” indicating the location of the news story

was not separately tagged in any of the GeoCLEF collection, but often appeared as the first part of the text for the story.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs used compounding in the indexing and querying processes to generate simple word forms from compounds.

The Snowball stemmer was used by both systems for language-specific stemming.

3.2 Search Processing

All of the runs for Monolingual English and German, and the runs for Bilingual English $\Rightarrow$ German used the standalone retrieval programs developed by Aitao Chen. The Monolingual Spanish and Portuguese, and the Bilingual English $\Rightarrow$ Spanish and English $\Rightarrow$ Portuguese runs all used the Cheshire II system.

The English and German Monolingual runs used language-specific compounding of German compound words. The Bilingual English $\Rightarrow$ German also used compounding.

Searching the GeoCLEF collection using the Cheshire II system involved using TCL scripts to parse the topics and submit the title and description or the title, description, and narrative from the topics. For monolingual search tasks we used the topics in the appropriate language (Spanish and Portuguese), for bilingual tasks the topics were translated from the source language to the target language using the L&H PC-based machine translation system. In all cases the various topic elements were combined into a single probabilistic query.

We tried two main approaches for searching, the first used only the topic text from the title and desc elements (TD), the second included the narrative elements as well (TDN). In all cases only the full-text “topic” index was used for Cheshire II searching.

Two of our English Monolingual runs used manual modification for topics 27, 43, and 50 by adding manually selected place names to the topics, in addition, one of these (which turned out to be our best performing English Monolingual run) also manually eliminated country names from topic 50.

After two initial runs for Portuguese Monolingual were submitted (BKGeoP1 and BKGeoP2), a revised and corrected version of the topics was released, and two additional runs (BKGeoP3 and BKGeoP4) were submitted using the revised topics, retaining the original submissions for comparison.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the submitted bilingual and monolingual runs for both English and German are shown in Table 1, the Recall-Precision curves for these runs are also shown in Figures 1 and 2 (for

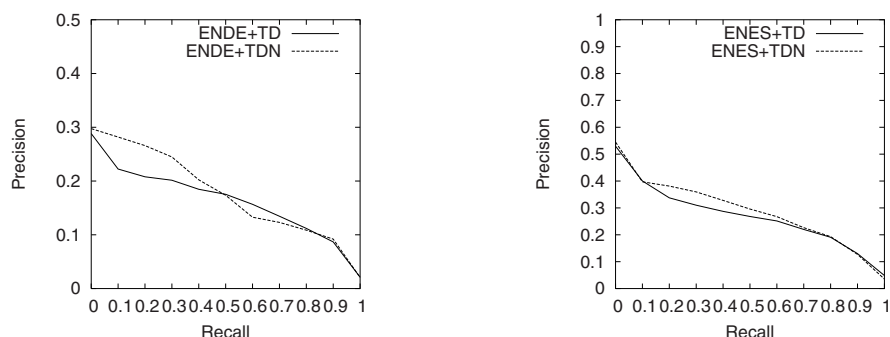


Fig. 3. Berkeley Bilingual Runs – English to German (left) and English to Spanish (right)

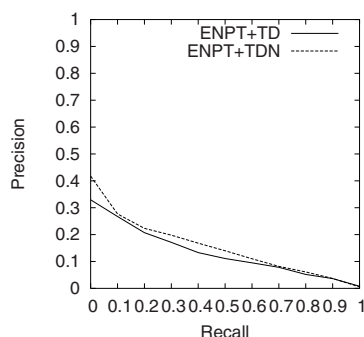


Fig. 4. Berkeley Bilingual Runs – English to Portuguese

monolingual) and 3 and 4 (for bilingual). In Figures 1-4 the names for the individual runs represent the language code and type of run, which can be compared with full names and descriptions in Table 1.

Table 1 indicates runs that had the highest overall MAP for the task by asterisks next to the run name. Single asterisks indicate the the highest MAP values among our own runs, while double asterisks indicate the runs where the MAP is the maximum recorded among official submissions.

As can be seen from the table, Berkeley’s cross-language submissions using titles, descriptions, and narratives from the topics were the best performing runs for the Bilingual tasks overall. Our Monolingual submissions, on the other hand did not fare as well, but still all ranked within the top quartile of results for each language except Portuguese where we fell below the mean. This result was surprising, given the good performance for Spanish. We now suspect that errors in mapping the topic encoding to the stored document encoding, or possibly problems with the Snowball stemmer for Portuguese may be responsible for this relatively poor performance.

Table 1. Submitted GeoCLEF Runs

Run Name	Description	Type	MAP
BKGeoED1	Bilingual English⇒German	TD auto	0.15612
BKGeoED2**	Bilingual English⇒German	TDN auto	0.16822
BKGeoES1	Bilingual English⇒Spanish	TD auto	0.25712
BKGeoES2**	Bilingual English⇒Spanish	TDN auto	0.27447
BKGeoEP1	Bilingual English⇒Portuguese	TD auto	0.12603
BKGeoEP2**	Bilingual English⇒Portuguese	TDN auto	0.14299
BKGeoD1*	Monolingual German	TD auto	0.21514
BKGeoD2	Monolingual German	TDN auto	0.18218
BKGeoE1	Monolingual English	TD auto	0.24991
BKGeoE2	Monolingual English	TDN auto	0.26559
BKGeoE3	Monolingual English	Manual	0.28268
BKGeoE4*	Monolingual English	Manual	0.28870
BKGeoS1*	Monolingual Spanish	TD auto	0.31822
BKGeoS2	Monolingual Spanish	TD auto	0.30032
BKGeoP1	Monolingual Portuguese	TD auto	0.16220
BKGeoP2	Monolingual Portuguese	TDN auto	0.16305
BKGeoP3	Monolingual Portuguese	TD auto	0.16925
BKGeoP4*	Monolingual Portuguese	TDN auto	0.17357

Table 2. Comparison of Berkeley’s best 2005 and 2006 runs for English and German

TASK	'06 NAME	MAP '06	Berk1 MAP '05	Berk2 MAP '05	Pct. Diff Berk1	Pct. Diff Berk2
GC-BILI-X2DE...	BKGeoED2	0.16822	0.0777	0.1137	116.50	47.95
GC-MONO-DE...	BKGeoD1	0.21514	0.0535	0.133	302.13	61.76
GC-MONO-EN...	BKGeoE4	0.28870	0.2924	0.3737	-1.26	-22.74

Last year’s GeoCLEF results (see [7]) also reported on runs using different systems (as Berkeley1 and Berkeley2), but both systems did all or most of the tasks. Table 2 shows a comparison of Average precision (MAP) for the best performing German and English runs for this year and for the two systems from last year. The German language performance of the system this year for both Bilingual and Monolingual tasks shows a definite improvement, while the English Monolingual performance is somewhat worse than either system last year. The “Berk2” system is essentially the same system as used this year for English and German runs.

5 Conclusions

Manual expansion of selected topics shows a clear, if small, improvement in performance over fully automatic methods. In comparing to Berkeley’s best performing English and German runs for last year, it would appear that either the

English queries this year were much more difficult, or that there were problems in the English runs. This year, while we did not use automatic expansion of toponyms in the topic texts, this was done explicitly in some of the topic narratives which may explain the improvements in runs using the narratives. It is also apparent that this kind of explicit toponym inclusion in queries, as might be expected, leads to better performance when compared to using titles and descriptions alone in retrieval.

Although we did not do any explicit geographic processing for this year, we plan to do so in the future. The challenge for next year is to be able to obtain the kind of effectiveness improvement seen with manual query expansion, in automatic queries using geographic processing.

References

1. Chen, A.: Multilingual information retrieval using english and Chinese queries. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 44–58. Springer, Heidelberg (2002)
2. Chen, A.: Cross-Language Retrieval Experiments at CLEF 2002. In: Peters, C., Braschler, M., Gonzalo, J. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 28–48. Springer, Heidelberg (2003)
3. Chen, A., Gey, F.C.: Multilingual information retrieval using machine translation, relevance feedback and compounding. *Information Retrieval* 7, 149–182 (2004)
4. Cooper, W.S., Chen, A., Gey, F.C.: Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In: Text REtrieval Conference (TREC-2), pp. 57–66 (1994)
5. Cooper, W.S., Gey, F.C., Dabney, D.P.: Probabilistic retrieval based on staged logistic regression. In: William, S., Cooper, F.C. (eds.) 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21–24, pp. 198–210. ACM Press, New York (1992)
6. Larson, R.R.: Probabilistic retrieval, component fusion and blind feedback for XML retrieval. In: Fuhr, N., Lalmas, M., Malik, S., Kazai, G. (eds.) INEX 2005. LNCS, vol. 3977, pp. 225–239. Springer, Heidelberg (2006)
7. Ray, R., Larson, F.C., Petras, V.: Berkeley at GeoCLEF: Logistic regression and fusion for geographic information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 963–976. Springer, Heidelberg (2006)
8. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms. *Journal of the American Society for Information Science*, 129–146 (1976)

UB at GeoCLEF 2006

Miguel E. Ruiz¹, June Abbas¹, David Mark², Stuart Shapiro³,
and Silvia B. Southwick¹

¹ State University of New York at Buffalo
Department of Library and Information Studies
534 Baldy Hall
Buffalo, NY 14260-1020 USA
meruiz@buffalo.edu
<http://www.informatics.buffalo.edu/faculty/ruiz>

² State University of New York at Buffalo
Department of Geography
105 Wilkeson Quad
Buffalo, NY 14261 USA

³ State University of New York at Buffalo
Department of Computer Science and Engineering
201 Bell Hall
Buffalo, New York 14260-2000 USA

Abstract. This paper summarizes the work done at the State University of New York at Buffalo (UB) in the GeoCLEF 2006 track. The approach presented uses pure IR techniques (indexing of single word terms as well as word bigrams, and automatic retrieval feedback) to try to improve retrieval performance of queries with geographical references. The main purpose of this work is to identify the strengths and shortcomings of this approach so that it serves as a basis for future development of a geographical reference extraction system. We submitted four runs to the monolingual English task, two automatic runs and two manual runs, using the title and description fields of the topics. Our official results are above the median system (auto=0.2344 MAP, manual=0.2445 MAP). We also present an unofficial run that uses title description and narrative which shows a 10% improvement in results with respect to our baseline runs. Our manual runs were prepared by creating a Boolean query based on the topic description and manually adding terms from geographical resources available on the web. Although the average performance of the manual run is comparable to the automatic runs, a query by query analysis shows significant differences among individual queries. In general, we got significant improvements (more than 10% average precision) in 8 of the 25 queries. However, we also noticed that 5 queries in the manual runs perform significantly below the automatic runs.

1 Introduction

For our participation in GeoCLEF 2006 we used pure information retrieval techniques to expand geographical terms present in the topics. We used a version

of the SMART[5] system that has been updated to handle modern weighting schemes (BM25, pivoted length normalization, etc.) as well as multilingual support (ISO-Latin1 encoding and stemming for 13 European languages using Porter's stemmer[3]).

We decided to work only with English documents and resources since they were readily available. Sections 2 and 3 present the details on collection preparation and query processing. Section 4 presents the retrieval model implemented with the SMART system. Section 5 shows results on the GeoCLEF 2005 data that was used for tuning parameters. Section 6 presents results using the official GeoCLEF 2006 topics as well as a brief analysis and discussion of the results. Section 7 presents our conclusions and future work.

2 Collection Preparation

Details about the GeoCLEF document collection are not discussed in this paper but the reader is referred to the GeoCLEF overview paper [1]. Our document collection consists of 169,477 documents from LA Times and The Glasgow Herald. Processing of English documents followed a standard IR approach discarding stop words and using Porter's stemmer. Additionally we added word bigrams that identify pairs of contiguous non-stop words to form a two word phrase. These bigrams allowed a stop word to be part of the bigram if they included the word "of" since it was identified as common component of geographical names (i.e. "United_Kingdom" and "City_of_Liverpool" would be valid bigrams). Documents were indexed using a vector space model (as implemented in the SMART system) with two ctypes. The first ctype was used to index words in the title and body of the article while the second ctype represented the indexing of the word bigrams previously described.

3 Query Processing

To process the topics we followed the same approach described above (using stop words, stemming, and adding word bigrams). Each query was represented using two ctypes. The first ctype for single word terms extracted from the parts that was used in the query (i.e. title and description). For our official runs we only used the title and description.

We designed a way to identify geographical features and expand them using geographical resources, but due to the short time available for developing the system, we could not include these automatically expanded terms in our official runs. For this reason we submitted results using a pure IR approach for this year and will work on the development of the geographical feature extraction for next year. Our results should be considered as baseline results.

Our manual runs were prepared by creating a Boolean query based on the topic description and manually adding terms from geographical resources available on the web. The following methodology was used to construct the manual queries:

1. An in-depth concept analysis was conducted on each query. Key concepts, geographic names, places, and features were identified.
2. Each key concept was then represented in a Boolean search string using AND, OR, and NOT operators.
3. Key concepts were expanded by the addition of synonyms and related terms when applicable.
4. Geographic names, places, and features web resources including National Geospatial Intelligence Agency (NGA) GEOnames Server¹, Geo-names.org database², and the Alexandria Digital Library (ADL) Gazetteer³ and ADL Feature Type Thesaurus⁴ were consulted.
5. For queries that included geographic radius elements, the ADL was used to determine additional place names to include in the query, based on specifications within the query and the GIS tools of the ADL. The ADL was also useful in determining cross-lingual names that were also included in the queries when necessary.

The manual run was included in the official results.

4 Retrieval Model

We used a generalized vector space model that combines the representation of the two ctypes and weighs the contribution of each part in the final similarity score between document and query. The final score is computed as the linear combination of ctype1 (words) and ctype2 (bigrams) as follows:

$$sim(d_i, q) = \lambda * sim_{words}(d_i, q) + \mu * sim_{bigrams}(d_i, q) \quad (1)$$

where λ and μ are coefficients that control the contribution of each of the two ctypes. The values of these coefficients were computed empirically using the optimal results in the GeoCLEF 2005 topics. The similarity values are computed using pivoted length normalization weighting scheme [6] ($pivot = 347.259$, $slope = 0.2$).

We also performed automatic retrieval feedback by retrieving 1000 documents using the original query and assuming that the top n documents are relevant and the bottom 100 documents are not relevant. This allowed us to select the top m terms ranked according to Rocchio's relevance feedback formula [4]:

$$w_{new}(t) = \alpha \times w_{orig}(t) + \beta \times \frac{\sum_{i \in Rel} w(t, d_i)}{\|Rel\|} - \gamma \times \frac{\sum_{i \in \neg Rel} w(t, d_i)}{\|\neg Rel\|} \quad (2)$$

where α , β , and γ are coefficients that control the contribution of the original query, the relevant documents (Rel) and the non-relevant documents ($\neg Rel$)

¹ <http://gnswww.nga.mil/geonames/GNS/index.jsp>

² <http://geonames.org>

³ <http://middleware.alexandria.ucsb.edu/client/gaz/adl/index.jsp>

⁴ <http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm>

respectively. The optimal values for these parameters are also determined using the GeoCLEF 2005 topics. Note that the automatic query expansion adds m terms to each of the two ctypes.

5 Preliminary Experiments Using CLEF2005 Topics

We first tested our baseline system using the GeoCLEF2005 topics. We used the title, description and geographic tags. Table 1 shows the performance values for the baseline run and for the best run submitted to GeoCLEF 2005 (BKGeoE1)[2]. The mean average precision for this baseline run is 0.3592 which is pretty good and would have been among the top three systems in GeoCLEF 2005. Figure 1 shows the recall-precision graph that compares the performance of our baseline, the retrieval feedback run and the best run in CLEF2005. We can see the figure shows very small differences in performance which indicates that a pure IR system was enough to answer most of the topics proposed last year.

Table 1. Performance of our baseline system against best run in GeoCLEF 2005

Parameters	UB Baseline	UB retrieval feedback	Best Run (BKGeoE1)
	$n = 5, m = 50$		
	$\alpha = 16, \beta = 96, \gamma = 8$		
	$\lambda = 10, \mu = 1$	$\lambda = 10, \mu = 1$	
MAP	36.42%	37.93%	39.36%
P@5	59.20%	58.40%	57.60%
P@10	46.80%	50.40%	48.00%
P@20	35.60%	37.20%	41.00%
P@100	18.16%	18.80%	21.48%

A query by query analysis revealed that the IR approach performed well in many topics but there are a few that could be improved. After analyzing these topics we conclude that most of them could have performed better if we had used some sort of expansion of continents using the countries located in them (i.e. European countries).

6 Results Using GeoCLEF 2006 Topics

We submitted four official runs, two using automatic query processing and two using manual methods. As expected our results (both automatic and manual) performed above the median system. Results are presented in Table 2.

The automatic runs perform slightly above the median system which indicates that the set of topics for this year were harder to solve using only IR techniques. After taking a look to the official topics we realize that we could have used a better expansion method using the geographical resources (i.e. identifying queries that have specific latitude and longitude references to restrict the set of retrieved results).

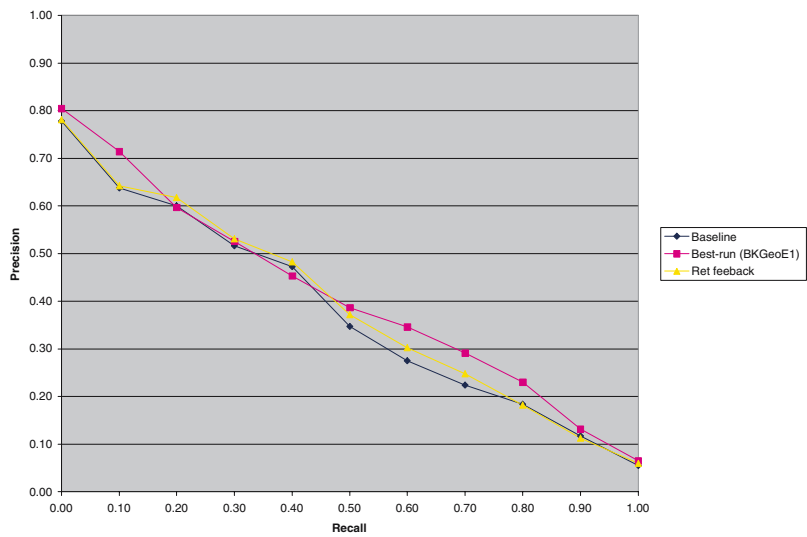


Fig. 1. Recall-Precision graph of our baseline and retrieval feedback systems against the best run in GeoCLEF 2005

Table 2. Performance of GeoCLEF 2006 Topics

Official Runs		
Run Label	Mean Avg. Precision Parameters	
UBGTDrf1 (automatic feedback)	0.2344	$n = 10, m = 20$ $\alpha = 16, \beta = 96, \gamma = 8,$ $\lambda = 10, \mu = 1$
UBGTDrf2 (automatic feedback)	0.2330	$n = 5, m = 50$ $\alpha = 16, \beta = 96, \gamma = 8,$ $\lambda = 10, \mu = 1$
UBGManual1 (Manual run only)	0.2307	$\lambda = 10, \mu = 1$
UBGManual2 (automatic feedback)	0.2446	$n = 10, m = 20$ $\alpha = 16, \beta = 96, \gamma = 8,$ $\lambda = 10, \mu = 1$
Unofficial Runs		
UBGTDRf1	0.2758	$n = 5, m = 5$ $\alpha = 16, \beta = 96, \gamma = 8,$ $\lambda = 10, \mu = 1$

On the other hand, the manual queries performed on average similarly to the automatic runs but a query by query analysis reveals that there are quite a few queries that significantly outperformed the automatic runs. However, at

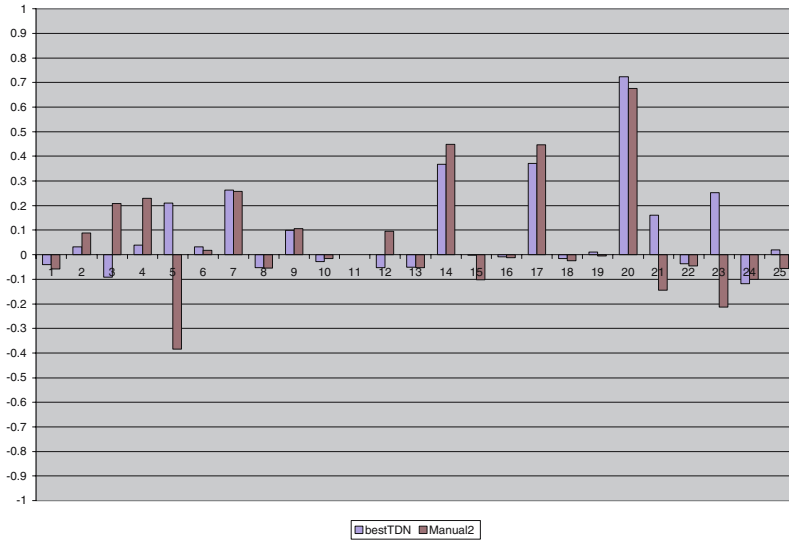


Fig. 2. Comparison of best manual run and best automatic run using our system

the same time there are two queries that performed significantly below the automatic systems. Note that the first manual run (UBGManual1) does not use automatic feedback while the second manual run (UBGManual2) uses automatic retrieval feedback. This merits further analysis to identify those strategies that are successful in improving performance.

Figure 2 compares the performance of our manual run and our best automatic run with respect to the median system in GeoCLEF. We noted that our best run (not submitted) performs quite well with respect to our baseline official runs. This run used title, description and narrative, and conservative retrieval feedback parameters ($n=5$ documents and $m=5$ terms). It is also encouraging that this run, when compared to the manual run, captured several of the good terms that were added manually.

Figure 3 shows a query by query comparison of the best manual run and the best automatic run obtained with our system and compares them against the median system in GeoCLEF 2006. Our main goal for this comparison is to try to identify patterns that could help in the future optimization of the system. After studying this graph we classified the queries according to the following criteria:

- Type 1: Queries where the automatic and manual runs performed significantly better than the median. (four queries: 32, 39, 42, 45)
- Type 2: Queries where the automatic system performed considerably below the median while the manual system performed significantly above. (two queries: 28 and 37)

- Type 3: Queries where the automatic and manual systems performed significantly below the median. (two queries: 33, 38)
- Type 4: Queries where the Automatic system performed significantly above the median and the manual system performed significantly below the median. (three queries: 30, 46, and 48)

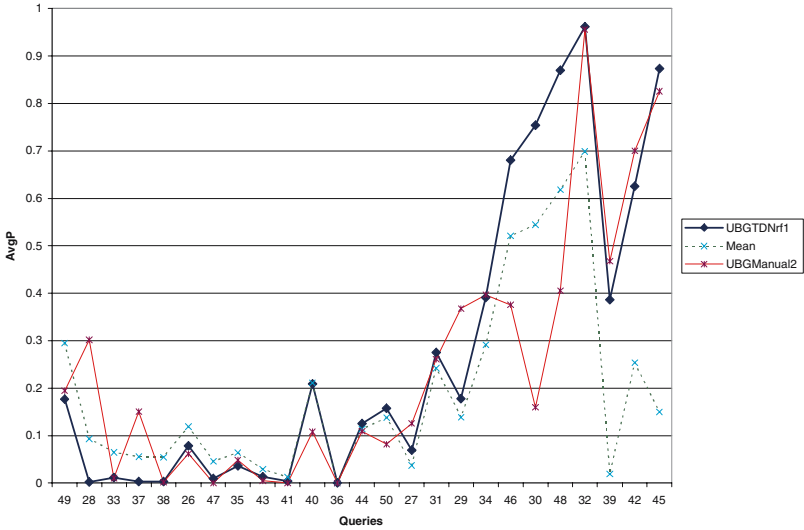


Fig. 3. Query by query comparison of best automatic run, best manual run and median system

From this analysis we learned that in most cases the major problem that affects negatively our performance seems to be linked to handling negation and exceptions (locations or events that are not considered relevant). On the other hand, word bigrams seem to be helping the retrieval of names and places that have two or more words. We also noticed that in some cases the improvement in performance is linked to usage of a non geographical term that plays an important role in determining the relevance of the related documents. For example, in query 28 the manual version uses the term blizzard as a synonym of snow storm. Also in this query the manual expansion includes geographical places where the snow storms are more common such as “Midwest United States”, “Northwest United States” and “Northeast United States”. At the same time, the manual query did not include any of the southern states that are warmer and snow fall is very rare. For an automatic system to capture this kind of details it should have some way to reason that blizzard is a synonym of snow storms and that they don’t occur in warm areas.

7 Conclusion

This paper presents an IR based approach to Geographical Information retrieval. Although this is our baseline system we can see that the results are competitive, especially if we use the long topics (title description and narrative). We still need to do more in depth analysis of the reasons why some manual queries improved significantly with respect to the median system and the problem presented in five queries that did perform significantly below the median. We plan to explore a way to generate automatic geographic references and ontology based expansion for next year.

Our query by query analysis shows some important patterns that can be used to guide our future development in building a more intelligent geographical terms expansion. Finally, our results show that even a simple approach that does not use sophisticated geographical term expansion can perform at a competitive level. This fact was observed with queries in GeoCLEF 2005 and 2006. This seems to indicate that the queries have a good deal of geographical clues and hence it would be hard to show that a sophisticated geographical term expansion would improve results.

References

- [1] Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Di Nunzio, G., Ferro, N.: Geoclef 2006: the clef 2006 cross-language geographic information retrieval track overview. In: Working Notes for the CLEF 2006 Workshop, Alicante, Spain (September 2006)
- [2] Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P.: Geoclef: The clef 2005 cross-language geographic information retrieval track overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 963–976. Springer, Heidelberg (2006)
- [3] Porter, M.F.: An algorithm for suffix stripping. *Program* 14, 130–137 (1980)
- [4] Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, Englewood Cliff, NJ (1971)
- [5] Salton, G.: *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs (1983)
- [6] Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, August 1996, pp. 21–29. ACM Press, New York (1996)

The University of Lisbon at GeoCLEF 2006

Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade,
and Mário J. Silva

Faculty of Sciences, University of Lisbon

{bmartins,ncardoso,mchaves,leonardo,mjs}@xldb.di.fc.ul.pt

Abstract. This paper details the participation of the XLDB Group from the University of Lisbon at the 2006 GeoCLEF task. We tested text mining methods that use an ontology to extract geographic references from text, assigning documents to encompassing geographic scopes. These scopes are used in document retrieval through a ranking function that combines BM25 text weighting with a similarity function for geographic scopes. We also tested a topic augmentation method, based on the geographic ontology, as an alternative to the scope-based approach.

1 Introduction

This paper reports the experiments of the XLDB Group from the University of Lisbon at the 2006 GeoCLEF task. Our main objective was to compare two specific strategies for Geographic IR, with a more standard IR approach. The specific strategies are: i) using text mining for extracting and combining geographic references from the texts, to assign documents to geographic scopes, together with a ranking function that combines scope similarity with a BM25 text ranking scheme, and ii) augmenting the geographic terms used in the topics through an ontology, again using a BM25 text ranking scheme to retrieve documents relevant to the augmented topics.

2 System Description

Figure 1 outlines the architecture of the GeoIR prototype used in our experiments. Many of the components came from a web search engine developed by our group, which is currently being extended with geographic IR functionalities (available at <http://local.tumba.pt>). For CLEF, the crawler was replaced by a simpler document loader, and the user interface was replaced by an interface specific for CLEF experiments that automatically parses CLEF topics and outputs results in the `trec_eval` format [1].

The components related to geographic text mining are shown in the gray boxes of Figure 1. They are assembled as a pipeline of operations for assigning documents to geographic scopes, and assigning topics to scopes. The system relies on an ontology that encodes place names and the semantic relationships among them, to recognize geographic terminology over documents and topics [2]. The CLEF topics are converted to *<what, relationship, where>* triples, where *what* is the non-geographical aspect, *where* is the geographic area of interest (i.e. geographic scope), and *relationship* is a spatial

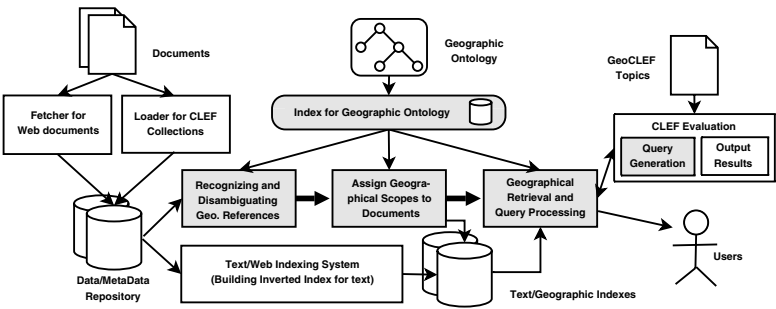


Fig. 1. Architecture of the geographical IR prototype assembled for GeoCLEF 2006

relationship connecting *what* and *where* [3]. The system’s final ranking combines a BM25 text weighting scheme [4] with a geographic scope similarity function.

For our text retrieval component, we set the BM25’s *k* and *b* parameters to the standard values of 2.0 and 0.75, and we used the extension proposed by Robertson et al. [4]. We also used a blind feedback expansion scheme for the topics. More details can be found in our paper of our participation at the CLEF-2006 adhoc task [1].

2.1 The Geographic Ontology

The ontology is a central component of the system, providing support for geographic reasoning. It models both the *vocabulary* and the *relationships* between geographic concepts in an hierarchical naming scheme with transitive “sub-region-of” and name alias capabilities. For our experiments, we developed an ontology with global geographic information in multiple languages, by integrating data from public sources [2]. Some statistics are listed in Figure 2. The information considered includes names for places and other geographic features, adjectives, place type information (e.g. street or city), relationships among concepts (e.g. adjacent or part-of), demographic data, spatial

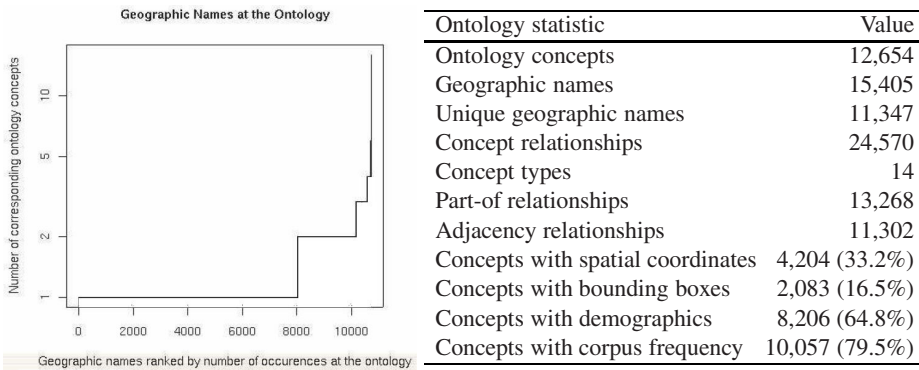


Fig. 2. Statistical characterization of the geographic ontology

coordinates (i.e. centroids) and bounding boxes for the geographic concepts. This ontology is freely available to all research groups at <http://xldb.di.fc.ul.pt/geonetpt>.

Each geographic concept can be described by several names. The chart presented on Figure 2 illustrates the ambiguity present in these names, by plotting the number of distinct associated concepts for each name. Even in our coarsely detailed ontology, place names with multiple occurrences are not just a theoretical problem (more than 25% of the place names correspond to multiple ontology concepts). As we do not have spatial coordinates or population information for some geographic concepts, we estimate those values from sibling concepts at the ontology (e.g. the centroid of a given region is approximated by the average of all centroids from its sub-regions, and the population of a region is the sum of the population for all its sub-regions). This heuristic is of particular importance, as we use these values on the geographic similarity function.

2.2 Recognizing Place References and Assigning Documents to Geographic Scopes

In the text mining approach, each document was assigned to a single encompassing geographic scope, according to the document's degree of locality. Each scope corresponds to a concept at our ontology. Scope assignment was performed off-line, as a pre-processing task that had two stages. First, we used a named entity recognition procedure, specifically tailored for recognizing and disambiguating geographic references in the text, relying on place names at the ontology and in lexical and contextual clues. Each reference was matched into the corresponding ontology concept (e.g. a string like "city of Lisbon" would be matched into a concept identifier at the ontology). Next, we combined the references extracted from each document into a single encompassing scope, using a previously described algorithm that explores relationships among geographic concepts [5]. The algorithm uses a graph-ranking approach to assign ontology concepts a confidence scores, selecting the highest scoring concept as the scope. For instance, if a document contains references to "Alicante" and "Madrid", it would be assigned to the scope "Spain", as both cities have a part-of relationship with that country.

On the CoNLL-03 English collection [6], our system has a precision of 0.85 and a recall of 0.79 for recognizing place references (reference disambiguation cannot be evaluated with this resource, as it lacks the associations from places to ontology concepts). The best reported system from CoNLL-03 achieved over 0.95 in both precision and recall, showing that our system can still be improved. As for the scope assignment procedure based on graph-ranking, it achieved an accuracy of 0.92 on the Reuters-21578 newswire collection [5].

2.3 Processing GeoCLEF Topics

GeoCLEF topics were also assigned to corresponding geographic scopes, so that we could match them to the scopes of the documents. Topic titles were converted into *<what, relationship, where>* triples, where *what* specifies the non-geographical aspect of the topic, *where* specifies the geographic area of interest (later disambiguated into a scope), and *relationship* specifies a spatial relation connecting *what* and *where* [3]. Two different types of relationships were found within the topics, namely near and

contained in. Topic GC40 (“cities near active volcanoes”) could not be processed by this mechanism, and was therefore treated as non-geographical (i.e. with no *where* and *relationship* terms). Some topics (e.g. GC29, “diamond trade in Angola and South Africa”) were assigned to multiple scopes, according to the different locations referenced in the *where* part.

2.4 Geographical Similarity

Geographic relevance ranking requires a mechanism for computing the similarity among the scopes assigned to the documents and the scopes assigned to the topics. Geographic scopes correspond to concepts at the ontology, and we can use the different sources of information, available at our ontology, to compute similarity. Based on previous results [7,8,9,10], we chose to use the following heuristics:

Topological Distance from Hierarchical Relationships. Topological part-of relationships can be used to infer similarity. We have, for instance, that “Alicante” is part of “Spain”, which in turn is part of “Europe”. “Alicante” should therefore be more similar with “Spain” than with “Europe”. We used the Equation (1), similar to Lin’s similarity measure [11], to compute the similarity according to the number of transitively common ancestors for the scope of a document and the scope of a topic.

$$OntSim(scope_{doc}, scope_{topic}) = \begin{cases} 1 & \text{if } scope_{doc} \text{ is the same or equivalent to } scope_{topic} \\ \frac{2 \times \text{NumCommonAncestors}(scope_{doc}, scope_{topic})}{\text{NumAncestors}(scope_{doc}) + \text{NumAncestors}(scope_{topic})} & \text{otherwise} \end{cases} \quad (1)$$

For example, considering the ontology on Figure 3, the similarity between the scopes corresponding to “Alicante” and “Spain” is $\simeq 0.67$, while the similarity between “Alicante” and “Europe” is 0.4.

Spatial Distance. Spatially close concepts are in principle more similar. However, people’s notion of distance depends on context, and one scope being near to another depends on the relative sizes and on the frame of reference.

We say that distance is zero, and therefore similarity is one, when one of the scopes is a sub-region of the other. We also normalize distance according to the diagonal of the minimum bounding rectangle for the scope of the topic, this way ensuring that different reference frames are treated appropriately. We employed a double sigmoid function with the center corresponding to the diagonal of the bounding rectangle. This function has a maximum value when the distance is at the minimum, and smoothly decays to zero as the distance increases, providing a non-linear normalization. The curve, represented in Figure 3, is given by Equation (2). D is the spatial distance between $scope_{doc}$ and $scope_{topic}$ and D_{MBR} is the diagonal distance of the minimum bounding rectangle for $scope_{topic}$. When $D = D_{MBR}$, the distance similarity is 0.5.

$$DistSim(scope_{doc}, scope_{topic}) = \begin{cases} 1 & \text{if } scope_{doc} \text{ is part-of or parent-of } scope_{topic} \\ 1 - \frac{1 + \text{sign}(D - D_{MBR}) \times (1 - \exp(-(\frac{D - D_{MBR}}{D_{MBR} \times 0.5})^2))}{2} & \text{otherwise} \end{cases} \quad (2)$$

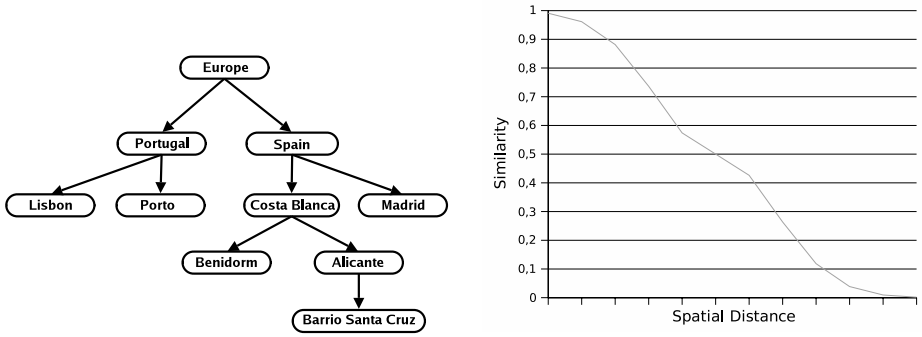


Fig. 3. An example ontology with hierarchical part-of relationships (left) and the double sigmoid function used to normalize the spatial distance (right)

Shared Population. When two regions are connected through a part-of relationship, the fraction of the population from the more general area that is also assigned to the more specific area can be used to compute a similarity measure. This metric corresponds to the relative importance of one region inside the other, and it also approximates the area of overlap. The general formula is given below:

$$PopSim(scope_{doc}, scope_{topic}) = \begin{cases} 1 & \text{if } scope_{doc} \text{ is the same or equivalent to } scope_{topic} \\ \frac{PopulationCount(scope_{doc})}{PopulationCount(scope_{topic})} & \text{if } scope_{doc} \text{ is part of } scope_{topic} \\ \frac{PopulationCount(scope_{topic})}{PopulationCount(scope_{doc})} & \text{if } scope_{topic} \text{ is part of } scope_{doc} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Adjacency from Ontology. Adjacent locations are, in principle, more similar than non-adjacent ones. Using the adjacency relationships from the ontology, we assign a score of 1 if the two scopes are adjacent, and 0 otherwise.

$$AdjSim(scope_{doc}, scope_{topic}) = \begin{cases} 1 & \text{if } scope_{doc} \text{ is adjacent to } scope_{topic} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

2.5 Score Combination for Geographic Retrieval and Ranking

The previously discussed measures, computed by different mechanisms, need to be combined into an overall similarity measure, accounting for textual and geographical aspects. We tried a linear combination due to its simplicity. Normalization is a crucial aspect to make different scores comparable. The previously given geographic measures already produce values in the interval $[0, 1]$. For the BM25 formula, we used the normalization procedure presented by Song et al. [12], corresponding to the formula below:

$$NormBM25(doc, topic) = \frac{\sum_{t_i \in doc} BM25(t_i) \times weight(topic, t_i)}{\sum_{t_i \in doc} \log\left(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5}\right)(k_1 + 1)} \quad (5)$$

The $weight(topic, t_i)$ parameter corresponds to 1 if term t_i is in the topic, and 0 otherwise. The final ranking score combined the normalized BM25 value with the similarity between the geographic scope of the document and the most similar scope of the topic (note that each topic could have more than one geographical scope assigned to the *where* term). It is given by the formula below:

$$Ranking(doc, topic) = (0.5 \times NormBM25(doc, topic)) + (0.5 \times MAX_{s \in scopes_{topic}}(GeoSim(scope_{doc}, s))) \quad (6)$$

where the geographical similarity $GeoSim$ is given by:

$$GeoSim(scope_{doc}, s) = (0.5 \times OntSim(scope_{doc}, s)) + (0.2 \times DistSim(scope_{doc}, s)) + (0.2 \times PopSim(scope_{doc}, s)) + (0.1 \times AdjSim(scope_{doc}, s)) \quad (7)$$

The combination parameters were based on the intuition that *topology matters and metric refines* [13], in the sense that we gave more weight to the similarity measures derived from topological relationships at the ontology. Still, for future work, we plan on using a systematic approach for finding the optimal combination. We also plan on using the confidence scores from the geographic scopes (recall that $scope_{doc}$ was assigned with a given confidence score) in ranking, adjusting the weight of $GeoSim$ accordingly.

3 Description of the Runs Submitted

Table 1 summarizes the submitted runs, a total of eight with half for the PT and half for the EN monolingual tasks. We did not submit runs for other languages, restricting our efforts to the PT and EN document collections.

In runs 2, 3, and 4, the non-geographical terms of each topic (i.e. the *where* terms obtained from the topic titles) were processed by the query expansion module. Essentially, the method adds the 15 top-ranked terms from the top 10 ranked documents of an initial ranking [1].

In run 3, ranked retrieval was based on the combination of BM25 with the similarity score computed between the scopes assigned to the topics and the scope of each document, as described in Section 2.5.

In run 4, the *where* terms were augmented, using information from the ontology to get related place names, either topologically or by proximity. A hierarchical structure can be used to expand place names in two directions, namely downward and upward [14]. Downward expansion is appropriate for queries involving a contained in

Table 1. Runs submitted to GeoCLEF 2006

Run Number	Description
1 (PT and EN)	Baseline using manually-generated queries from the topics and BM25 text retrieval.
2 (PT and EN)	BM25 text retrieval. Queries were generated from query expansion of <i>what</i> terms at the topic title, together with the original <i>where</i> and <i>relationship</i> terms also at the topic title.
3 (PT and EN)	Geographic relevance ranking using geographic scopes. Queries were generated from query expansion of <i>what</i> terms at the topic title. The <i>where</i> terms in the topic title were matched into scopes.
4 (PT and EN)	BM25 text retrieval. Queries were generated from query expansion of <i>what</i> terms at the topic title, together with the augmentation of <i>where</i> and <i>relationship</i> terms also at the topic title.

spatial relationship, extending the influence of a place name to all of its descendants, in order to encompass sub-regions of the location specified in the query. Upward expansion can be used to extend the influence of a place name to some or all of its ancestors, and then possibly downward again into other sibling places. This can be used for queries involving a near spatial relation, although many irrelevant place names can this way also be included. We have chosen not to use upwards expansion, instead using adjacency relationships from the ontology and near concepts computed from the spatial coordinates. The general augmentation procedure involved the following steps:

1. Use the ontology to get concepts that correspond to sub-regions of the *where* term(s) obtained from the topic title (i.e. topologically related concepts).
2. If the *relationship* term obtained from the topic title corresponds to the near relationship, use the ontology to get the adjacent regions to the *where* term(s).
3. If the *relationship* term obtained from the topic title corresponds to the near relationship, use the ontology to get the top k nearest locations from the *where* term(s).
4. Rank the list of concepts that was obtained from the previous three steps according to an operational notion of importance. This ranking procedure is detailed in a separate publication [3], and considers heuristics such as concept types (e.g. countries are preferred to cities, which in turn are preferred to small villages), demographics, and occurrence frequency statistics for the place names.
5. Select the place names from the 10 top ranked concepts to augment the original topic.

4 Results

Table 2 summarizes the `trec_eval` output for the official runs. Run 1 achieved the best results, both in the PT and EN subtasks, obtaining MAP scores of 0.301 and 0.303, respectively. Contrary to our expectations, run 4 also outperformed run 3, showing that a relatively simple augmentation scheme for the geographic terminology of the topics can outperform the text mining approach. In GeoCLEF 2005, our best run achieved a MAP score of 0.2253 (also a baseline with manually-generated queries). Similarly, our run

Table 2. Results obtained for the different submitted runs

Measure	Run 1		Run 2		Run 3		Run 4	
	PT	EN	PT	EN	PT	EN	PT	EN
num-q	25	25	25	25	25	25	25	25
num-ret	5232	3324	23350	22483	22617	21228	10483	10652
num-rel	1060	378	1060	378	1060	378	1060	378
num-rel-ret	607	192	828	300	519	240	624	260
map	0,301	0,303	0,257	0,158	0,193	0,208	0,293	0,215
R-prec	0,359	0,336	0,281	0,153	0,239	0,215	0,346	0,220
P5	0,488	0,384	0,416	0,208	0,432	0,240	0,536	0,288
P10	0,496	0,296	0,392	0,180	0,372	0,228	0,480	0,240
P20	0,442	0,224	0,350	0,156	0,318	0,170	0,424	0,212
P100	0,218	0,072	0,193	0,073	0,162	0,068	0,218	0,084

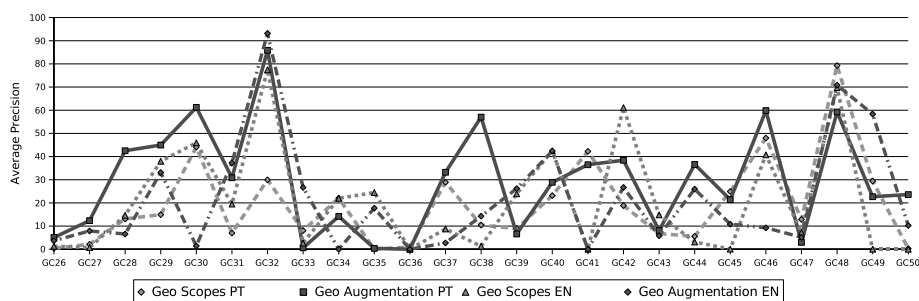


Fig. 4. Average precision for the 25 topics in runs 3 and 4, for both PT and EN subtasks

submissions to GeoCLEF 2005 used an automatic technique that involved geographic scope assignment. However, the retrieval scheme was less elaborate, and we achieved a MAP score of 0.1379 [15]. Figure 4 shows the average precision for the 25 individual topics, for runs 3 and 4 and in the PT and EN subtasks.

We analysed the documents retrieved for some of the topics, together with the scopes that had been assigned to them, particularly focusing on GC32 and GC48. We believe that run 3 performed worse due to errors in scope assignment and to having each document assigned to a single geographic scope, which seems too restrictive. We are now performing additional experiments, using the GeoCLEF 2006 relevance judgments, by reassigning geographic scopes to the documents and this time allowing the association of multiple scopes to each document.

5 Conclusions

We mainly tested two different approaches at GeoCLEF 2006: i) the relatively simple augmentation of geographic terms in the topics, through the use of a geographic ontology; and ii) a text mining approach based on extracting geographical references from documents, in order to assign a geographic scope to each document. In the latter approach, relevance ranking was based on a linear combination of the BM25 text weighting scheme with a similarity function for scopes. In both cases, the obtained results were of acceptable quality, although they could not meet the expectations. Particularly, the text mining approach failed in providing better results than the augmentation method. This point requires more investigation, and we are already making additional experiments with the relevance judgments for the GeoCLEF 2006 topics [16].

Acknowledgements

We thank Daniel Gomes who built the tumba! repository. Our participation was partially supported by grants POSI/PLP/43931/2001 (Linguatca), SFRH/BD/10757/2002 and POSI/SRI/40193/2001 (GREASE) from FCT (Portugal), co-financed by POSI.

References

1. Cardoso, N., Silva, M.J., Martins, B.: The University of Lisbon at CLEF 2006 Ad-Hoc Task. In: Working notes for the CLEF 2006 Workshop (2006)
2. Chaves, M., Martins, B., Silva, M.J.: GKB - Geographic Knowledge Base. DI/FCUL TR 05-12, Department of Informatics, University of Lisbon (2005)
3. Martins, B., Silva, M.J., Freitas, S., Afonso, A.P.: Handling Locations in Search Engine Queries. In: Proceedings of the 3rd Workshop on Geographical IR, GIR-06 (2006)
4. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields. In: Proceedings of the 13th International Conference on Information and Knowledge Management, CIKM-04, ACM Press, New York (2004)
5. Martins, B., Silva, M.J.: A Graph-Ranking Algorithm for Geo-Referencing Documents. In: Proceedings of the 5th International Conference on Data Mining, ICDM-05 (2005)
6. Sang, F.T.K., F., E., Meulder, D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003 (2003)
7. Alani, H., Jones, C.B., Tudhope, D.: Associative and Spatial Relationships in Thesaurus-Based Retrieval. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, Springer, Heidelberg (2000)
8. Gutiérrez, M., Rodríguez, A.: Querying Heterogeneous Spatial Databases: Combining an Ontology with Similarity Functions. In: Proceedings of the 1st International Workshop on Conceptual Modeling for GIS, CoMoGIS-2004 (2004)
9. Jones, C.B., Alani, H., Tudhope, D.: Geographical Information Retrieval with Ontologies of Place. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, Springer, Heidelberg (2001)
10. Rodríguez, A., Egenhofer, M.: Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographic Information Science* 18 (2004)
11. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proceedings of the 15th International Conference on Machine Learning, ICML-98 (1998)
12. Song, R., Ji-RongWen, Shi, S., Xin, G., Tie-YanLiu, Qin, T., Xin Zheng, J.Z., Xue, G., Ma, W.Y.: Microsoft Research Asia at the Web Track and TeraByte Track of TREC 2004. In: Proceedings of the 13th Text REtrieval Conference, TREC-04 (2004)
13. Egenhofer, M.J., Mark, D.M.: Naïve Geography. In: Kuhn, W., Frank, A.U. (eds.) COSIT 1995. LNCS, vol. 988, Springer, Heidelberg (1995)
14. Li, Y., Moffat, A., Stokes, N., Cavedon, L.: Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In: Proceedings of the 3rd Workshop on Geographical IR, GIR-06 (2006)
15. Cardoso, N., Martins, B., Chaves, M., Andrade, L., Silva, M.J.: The XLDB Group at Geo-CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
16. Andrade, L., Silva, M.J.: Relevance Ranking for Geographic IR. In: Proceedings of the 3rd Workshop on Geographical Information Retrieval, GIR-2006 (2006)

Author Index

- Abbas, June 978
 Abdou, Samir 137
 Adriani, Mirna 57, 848
 Agosti, Maristella 11
 Ahn, David 362
 Al-Maskari, Azzah 205
 Alexandrov-Kabadjov, Mijail 541
 Alvarado, X. 450
 Aly, Robin 770
 Alzghool, Muath 778
 Andrade, Leonardo 986
 Arcoverde, João Marcelo Azevedo 74
 Argaw, Atelach Alemu 43
 Ariza-López, F. Javier 918
 Ariles, Javier 195
 Asker, Lars 43
 Ayache, Christelle 223
 Azzopardi, Leif 803

 Baerisch, Stefan 163
 Bagur, Michael 454
 Balage Filho, Pedro Paulo 372
 Balog, Krisztian 803, 830
 Barbu, Eugen 670
 Beigbeder, Michel 83
 Bellot, Patrice 440, 889
 Bensrhair, Abdelaziz 670
 Berrocal, José L. Alonso 145, 820
 Besançon, Romaric 395
 Bhattacharya, P. 715
 Bischoff, Kerstin 852, 946
 Blaudez, Eric 440, 889
 Bos, Johan 290
 Bosma, Wauter 502
 Bouma, Gosse 318
 Bowden, Mitchell 310
 Buriol, Luciana S. 91
 Burkhardt, Hans 644
 Buscaldi, Davide 377, 550, 918, 954

 Callison-Burch, Chris 502
 Cardoso, Nuno 51, 986
 Cassan, Adán 300
 Cavedon, Lawrence 938

 Ceașu, Alin 385
 Chang, Yih-Chen 625
 Chatain, P. 351
 Chaves, Marcirio Silveira 986
 Chen, Hsin-Hsi 625
 Chevallet, Jean-Pierre 694, 735
 Clark, Jonathan 310
 Clough, Paul 186, 205, 579, 595
 Coelho, Alexandre Ramos 91
 Coppola, Bonaventura 526
 Costa, Luís 405, 569
 Cristea, Dan 385

 Darmoni, Stefan 670
 Darwish, Kareem 205
 Datta, Kalyankumar 107
 Daumke, Philipp 652
 de Jong, Franciska 770
 de Lima, Vera Lúcia Strube 66
 de Pablo-Sánchez, César 463
 de Rijke, Maarten 265, 362, 537, 803, 830
 de Uzêda, Vinícius Rodrigues 372
 Denicia-Carral, Claudia 415
 Desai, B.C. 715
 Deselaers, Thomas 579, 595, 725
 Deserno, Thomas M. 595, 686
 Desmontils, E. 280
 Díaz-Galiano, M.C. 711
 Di Nunzio, Giorgio Maria 11, 21, 852
 Dixon, P.R. 351
 Dornescu, Iustin 385

 Eibl, Maximilian 178, 739
 El-Bèze, Marc 440, 889
 Embarek, Mehdi 395

 Fahmi, Ismail 318
 Ferrández, A. 450
 Ferrández, Oscar 490, 889, 918
 Ferrández, S. 450
 Ferrés, Daniel 962
 Ferret, Olivier 395
 Ferro, Nicola 11, 21, 852
 Figueira, Helena 300

- Figuerola, Carlos G. 145, 820
 Fischer, Benedikt 686
 Fissaha Adafre, Sisay 537
 Florea, Filip 670
 Forner, Pamela 223
 Furui, S. 351

 García-Cumbreras, Miguel Á. 119, 711, 913, 918
 García-Serrano, Ana M. 613
 García-Vega, Manuel 913, 918
 Gass, Tobias 638
 Ge, Linlin 905
 Geissbuhler, Antoine 638
 Geoffrey, Andogah 881
 Gey, Fredric C. 852, 970
 Giampiccolo, Danilo 223
 Gillard, Laurent 440, 889
 Glöckner, Ingo 518
 Gobeill, Julien 706
 Gomez, José Manuel 377
 Goñi-Menoyo, José M. 877
 González-Cristóbal, José C. 877
 González-Ledesma, Ana 463
 Gonzalez, Marco 66
 Gonzalo, Julio 186, 195
 Grau, Brigitte 454
 Grubinger, Michael 579
 Guillén, Rocio 893
 Güld, Mark O. 686

 Hackl, René 127
 Halácsy, Péter 99
 Halawani, Alaa 644
 Hanbury, Allan 579
 Hartrumpf, Sven 432
 Hauff, Claudia 958
 Hayurani, Herika 57
 Heie, M.H. 351
 Herrera, Jesús 483
 Hersh, William 595, 660
 Heuwing, Ben 834
 Hiemstra, Djoerd 770
 Hu, Yang 678
 Hu, You-Heng 905
 Huang, Xiaoli 744

 Iftene, Adrian 385
 Iles, Brandon 494
 Inkpen, Diana 778
 Inoue, Masashi 617

 Ion, Radu 385
 Ircing, Pavel 759
 Izquierdo Beviá, Rubén 554

 Jacquin, C. 280
 Jensen, Jeffery 660
 Jijkoun, Valentin 223, 265, 362, 537
 Jiménez, Ernesto 826
 Jiménez-Salazar, Héctor 838
 Jones, Gareth J.F. 153, 633, 744, 794
 Juárez-González, Antonio 415

 Kalpathy-Cramer, Jayashree 660
 Kamps, Jaap 803
 Karlgren, Jussi 186, 217
 Khoerniawan, Tommy 848
 Kim, Eugene 595
 Kouylekov, Milen 526
 Kozareva, Zornitsa 522, 889, 918
 Kruschwitz, Udo 541
 Kürsten, Jens 178

 Lacoste, Caroline 694
 Lam-Adesina, Adenike M. 153, 794
 Lana-Serrano, Sara 877
 Larson, Ray R. 174, 609, 852, 970
 Laurent, Dominique 339
 Le, Diem Thi Hoang 694
 Leveling, Johannes 170, 432, 901
 Li, Mingjing 678
 Li, Yi 938
 Li, Zhisheng 926
 Ligozat, Anne-Laure 454
 Lim, Joo-Hwee 694, 735
 Liu, Jing 678
 Llopis, Fernando 62, 450
 López-López, Aurelio 424
 López-Moreno, P. 450
 López-Ostenero, Fernando 195

 Ma, Songde 678
 Ma, Wei-Ying 678, 926
 Magalhães, João 930
 Magnini, Bernardo 223, 526
 Maillot, Nicolas 735
 Majumder, Prasenjit 107
 Mandl, Thomas 21, 127, 834, 852, 946
 Mark, David 978
 Marko, Kornel 652
 Martín-Valdivia, M.T. 711

- Martínez, Paloma 613
 Martínez-Barco, Patricio 490, 766
 Martínez-Fernández, J.L. 613
 Martínez-González, Ángel 877
 Martínez-Santiago, Fernando 119
 Martins, André 300
 Martins, Bruno 51, 986
 McDonald, Kieran 633
 McNamee, Paul 157
 Mendes, Afonso 300
 Mendes, Pedro 300
 Mercier, Annabelle 83
 Mitra, Mandar 107
 Moffat, Alistair 938
 Moldovan, Dan 310, 494
 Monceaux, L. 280
 Montejo-Ráez, Arturo 119, 711
 Montes-y-Gómez, Manuel 415, 424
 Montoyo, Andrés 522, 889, 918
 Moreno Monteagudo, Lorenza 554
 Moreno-Sandoval, Antonio 463
 Moruz, Alex 385
 Moschitti, Alessandro 510
 Müller, Henning 579, 595, 638, 706
 Müller, Luděk 759
 Muñoz, Rafael 490, 889, 918
 Mur, Jori 318

 Nègre, Sophie 339
 Negri, Matteo 526
 Neumann, Günter 328
 Ney, Hermann 725
 Nissim, Malvina 290
 Noguera, Elisa 62, 450, 889, 918
 Novak, J.R. 351
 Nunes, Maria das Graças Volpe 74, 372

 Oakes, Michael P. 111
 Oard, Douglas W. 744, 786
 Olsson, Fredrik 217
 Olteanu, Marian 310
 Orăsan, Constantin 561
 Ordelman, Roeland 770
 Orenge, Viviane Moreira 91
 Osenova, Petya 223
 Overell, Simon 930

 Paetzold, Jan 652
 Palomar, Manuel 490, 766
 Pancardo-Rodríguez, Aarón 424

 Pardo, Thiago Alexandre Salgueiro 372
 Pecina, Pavel 744
 Peinado, Víctor 195
 Peñas, Anselmo 223, 257, 483
 Peral, J. 450
 Perea-Ortega, José M. 913, 918
 Pérez-Coutiño, Manuel 424
 Peters, Carol 1, 21
 Pingali, Prasad 35
 Pinto, Cláudia 300
 Pinto, David 826, 838
 Pistol, Ionuț 385
 Poesio, Massimo 541
 Pușcașu, Georgiana 385, 554, 561

 Qiu, Bo 690

 Racocanu, Daniel 694
 Rahman, M.M. 715
 Robba, Isabelle 454
 Rocha, Paulo 223, 852
 Rode, Henning 958
 Rodrigo, Álvaro 257, 483
 Rodríguez, Emilio 820
 Rodríguez, Horacio 962
 Roger, S. 450
 Rogozan, Alexandrina 670
 Rojas, Franco 838
 Rosso, Paolo 377, 550, 826, 918, 954
 Ruch, Patrick 706
 Rüger, Stefan 930
 Ruiz, Miguel E. 702, 978

 Sacaleanu, Bogdan 223, 328
 Sama, Valentín 257
 Sanchis, Emilio 377, 954
 Sanderson, Mark 852
 Santos, Diana 569, 852
 Saquete Boró, Estela 554
 Sari, Syandra 57
 Sarmento, Luís 473
 Savoy, Jacques 137
 Séguéla, Patrick 339
 Séjourné, Kevin 454
 Setia, Lokesh 644
 Shafran, Izhak 744
 Shapiro, Stuart 978
 Silva, Mário J. 51, 986
 Sitbon, Laurianne 440, 889
 Soergel, Dagobert 744

- Sood, V. 715
 Southwick, Silvia B. 978
 Ștefănescu, Dan 385
 Steinberger, Josef 541
 Stempfhuber, Maximilian 163
 Stokes, Nicola 938
 Strötgen, Robert 834
 Suriyentrakor, Pasin 310
 Sutcliffe, Richard F.E. 223, 541

 Tait, John I. 111
 Tatu, Marta 494
 Téllez-Valero, Alberto 415
 Teodorescu, Roxana 694
 Terol, Rafael M. 490, 766
 Teynor, Alexandra 644
 Thies, Christian 686
 Tiedemann, Jörg 318
 Tjong Kim Sang, Erik 362
 Tomás, David 275
 Tomlinson, Stephen 129, 844
 Toral Ruiz, Antonio 554, 889, 918
 Trandabăt, Diana 385
 Trieschnigg, Dolf 958
 Trón, Viktor 99
 Tufiş, Dan 385
 Tune, Kula Kekeba 35

 Ureña-López, L. Alfonso 119, 711,
 913, 918

 Vázquez, Sonia 522
 van der Plas, Lonneke 318

 van der Werff, Laurens 770
 van Noord, Gertjan 318
 van Rantwijk, Joris 362
 Varma, Vasudeva 35
 Veiel, Dirk 901
 Verdejo, Felisa 257, 483
 Vicedo, José L. 275
 Vicente-Díez, Maria Teresa 463
 Vidal, Daniel 300
 Vilnat, Anne 454
 Vilares, Jesús 111
 Villaseñor-Pineda, Luis 415, 424
 Villena Román, Julio 613
 Vuillenemot, Nicolas 694

 Wang, Chong 926
 Wang, Jianqiang 744, 786
 Wang, Xufa 926
 Weyand, Tobias 725
 White, Ryen W. 744
 Whittaker, E.W.D. 351
 Widhi, Bimo 848
 Wijaya, Syntia 848
 Wilhelm, Thomas 739
 Womser-Hacker, Christa 127, 852, 946

 Xie, Xing 926
 Xiong, Wei 694

 Zanzotto, Fabio Massimo 510
 Zazo, Angel F. 145
 Zazo Rodríguez, Ángel F. 820
 Zhang, Ke 794