

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

Alfred Kobsa

University of California, Irvine, CA, 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

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Carol Peters Valentin Jijkoun
Thomas Mandl Henning Müller
Douglas W. Oard Anselmo Peñas
Vivien Petras Diana Santos (Eds.)

Advances in Multilingual and Multimodal Information Retrieval

8th Workshop of the Cross-Language Evaluation Forum,
CLEF 2007
Budapest, Hungary, September 19-21, 2007
Revised Selected Papers

Volume Editors

Carol Peters

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

Valentin Jijkoun

University of Amsterdam, The Netherlands; jijkoun@science.uva.nl

Thomas Mandl

University of Hildesheim, Germany; mandl@uni-hildesheim.de

Henning Müller

University of Applied Sciences Western Switzerland, Sierre, Switzerland and
University Hospitals and University of Geneva, Switzerland
henning.mueller@sim.hcuge.ch

Douglas W. Oard

University of Maryland, USA; oard@glue.umd.edu

Anselmo Peñas

Universidad Nacional de Educación a Distancia, Madrid, Spain; anselmo@lsi.uned.es

Vivien Petras

GESIS Social Science Information Centre, Bonn, Germany; vivien.petras@gesis.org

Diana Santos

Linguatca, SINTEF, Norway; diana.santos@sintef.no

Managing Editor: Danilo Giampiccolo, CELCT, Trento, Italy; giampiccolo@celct.it

Library of Congress Control Number: 2008934678

CR Subject Classification (1998): I.2.7, H.5, I.2, H.2, I.7, H.4

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

ISSN 0302-9743

ISBN-10 3-540-85759-1 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-85759-4 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 2008

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12465019 06/3180 5 4 3 2 1 0

Preface

The eighth campaign of the Cross Language Evaluation Forum (CLEF) for European languages was held from January to September 2007. There were seven distinct evaluation tracks in CLEF 2007, designed to test the performance of a wide range of multilingual information access systems or system components. CLEF is by now an established international evaluation initiative and, in 2007, 81 groups from all over the world submitted results for one or more of the different evaluation tracks. Full details regarding the design of the tracks, the methodologies used for evaluation, and the results obtained by the participants can be found in the different sections of these proceedings.

As always the results of the campaign were reported and discussed at the annual workshop, held in Budapest, Hungary, 19-21 September, immediately following the eleventh European Conference on Digital Libraries. The workshop plays an important role by providing the opportunity for all the groups that have participated in the evaluation campaign to get together to compare approaches and exchange ideas.

The schedule of the workshop was divided between plenary track overviews and parallel, poster and breakout sessions presenting the 2007 experiments and discussing ideas for the future. There were also several invited talks. Noriko Kando, National Institute of Informatics, Tokyo, reported the lessons learned at NTCIR-6 and plans for NTCIR-7 (NTCIR is an evaluation initiative focussed on testing IR systems for Asian languages), while Mandar Mitra, Indian Statistical Institute Kolkata, presented FIRE, a new Forum for Information Retrieval Evaluation for Indian languages. Edouard Geoffrois of the French government described the objectives of the ambitious Quaero programme, which has the goal of developing multimedia and multilingual indexing and management tools for professional and general public applications. In two final talks, Martin Braschler of the Zurich University of Applied Sciences gave a summing up of what he felt were the major trends of the 2007 campaign in the light of what had emerged during the discussions at the workshop, and Donna Harman of the US National Institute of Standards and Technology made a number of recommendations for the future. In particular, she urged the participants to focus more on failure analysis; not just to recognise what methods were working and what had little impact but to try to really understand why and then to think about generalizing what has been learnt. The presentations given at the CLEF workshop can be found on the CLEF website at www.clef-campaign.org.

CLEF 2007 was an activity of the DELOS Network of Excellence for Digital Libraries, of the Information Society Technologies programme of the European Commission. However, synergy of activities has been established with other Networks. The CLEF workshop was thus preceded by two related events. On September 18, the ImageCLEF group, together with the MUSCLE Network

of Excellence held a joint meeting on the “Evaluation of Image and Video Retrieval”. On the morning of September 19, the MorphoChallenge 2007 meeting on “Unsupervised Morpheme Analysis” was held. MorphoChallenge 2007 was part of the EU Network of Excellence PASCAL Challenge Program and was organized in collaboration with CLEF.

At the time of writing, the organisation of CLEF 2008 is well underway. The campaign this year again includes seven main evaluation tracks. Pilot tasks are also proposed to assess the performance of systems working on cross-language video retrieval and multilingual information filtering. In addition, for the first time, in 2008 CLEF offers testing with target collections in non-European languages: Arabic in the filtering track and Persian for Ad Hoc retrieval tasks. This is an important new step for CLEF; the implications will be discussed in the next workshop, to be held in Aarhus, Denmark, September 17–19, 2008. CLEF 2008 is sponsored by TrebleCLEF, a Coordination Action of the Seventh Framework Programme of the European Commission.

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 final 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.

May 2008

Carol Peters
Valentin Jijkoun
Thomas Mandl
Henning Müller
Douglas W. Oard
Anselmo Peñas
Vivien Petras
Diana Santos

Reviewers

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

- Mirna Adriani, Faculty of Computer Science, University of Indonesia, Indonesia
- Stefan Baerisch, GESIS Social Science Information Centre, Bonn, Germany
- Paul D. Clough, University of Sheffield, UK
- Thomas Deselaers, RWTH Aachen University, Germany
- Thomas M. Deserno, RWTH Aachen University, Germany
- Giorgio Di Nunzio, University of Padua, Italy
- Nicola Ferro, University of Padua, Italy
- Corina Forascu, Institute for Research in Artificial Intelligence, Romania
- Fredric C. Gey, University of California at Berkeley, USA
- Ingo Glöckner, FernUniversität in Hagen, Germany
- Michael Grubinger, Victoria University, Melbourne, Australia
- Danilo Giampiccolo, Centre for the Evaluation of Human Language and Multimodal Communication Technologies (CELCT), Trento, Italy
- Rene Hackl-Sommer, University of Hildesheim and Fachinformationszentrum Karlsruhe, Germany
- Allan Hanbury, Technical University of Vienna, Austria
- Donna Harman, National Institute of Standards and Technology, USA
- Sven Hartrumpf, FernUniversität in Hagen, Germany
- William Hersh, Oregon Health and Science University, USA
- Diana Inkpen, University of Ottawa, Canada
- Gareth Jones, Dublin City University, Ireland
- Jayashree Kalpathy-Cramer, Oregon Health and Science University, USA
- Ralph Kölle, University of Hildesheim, Germany
- Mikko Kurimo, Helsinki University of Technology, Finland
- Ray Larson, University of California at Berkeley, USA
- Matthew Lease, Brown University, USA
- Johannes Leveling, FernUniversität in Hagen, Germany
- Paul McNamee, Johns Hopkins University, USA
- Diego Molla, Macquarie University, Australia
- Manuel Montes, INAOE, Mexico
- Günter Neumann, German Research Centre for Artificial Intelligence, Germany
- Petya Osenova, Bulgarian Academy of Sciences
- Ross Purves, University of Zürich - Irchel, Switzerland
- Paulo Rocha, Linguatca, Sintef, Norway
- Alvaro Rodrigo, UNED, Madrid, Spain
- Paolo Rosso, Polytechnic University of Valencia, Spain

- Miguel Ruiz, University of North Texas, USA
- Mark Sanderson, University of Sheffield, UK
- Jacques Savoy, University of Neuchatel, Switzerland
- Mário J. Silva, University of Lisbon, Portugal
- Stephen Tomlinson, Open Text Corporation, Canada
- Jordi Turmo, Polytechnic of Catalonia, Spain
- José-Luis Vicedo, University of Alicante, Spain
- Wouter Weerkamp, University of Amsterdam, The Netherlands
- Christa Womser-Hacker, University of Hildesheim, Germany
- Xing Xie, Microsoft Research Asia, China
- Fabio Massimo Zanzotto, University of Rome “Tor Vergata”, Italy

CLEF 2007 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 2007 campaign:

- Centre for the Evaluation of Human Language and Multimodal Communication Technologies (CELCT), Italy
- College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, USA
- Department of Computer Science, University of Indonesia, Indonesia
- Department of Computer Science and Information Systems, University of Limerick, Ireland
- Department of Computer Science and Information Engineering, National University of Taiwan, 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 and Computer Science, RWTH Aachen University, Germany
- Evaluations and Language Resources Distribution Agency (ELDA), France
- Fondazione Bruno Kessler FBK-irst, Trento, Italy
- German Research Centre 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
- Institute of Formal and Applied Linguistics, Charles University, Czech Rep.
- Universidad Nacional de Educación a Distancia, Madrid, Spain
- Linguatca, Sintef, Oslo, Norway
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, Bulgaria
- Microsoft Research Asia
- National Institute of Standards and Technology, Gaithersburg, USA
- Department of Medical Informatics & Clinical Epidemiology, 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 Computing, Dublin City University, Ireland
- School of Computer Science and Mathematics, Victoria Univ., Australia
- School of Information Management and Systems, UC Berkeley, USA
- Universitat Politècnica de Catalunya, Barcelona, Spain
- University “Alexandru Ioan Cuza”, IASI, Romania
- University and University Hospitals of Geneva, Switzerland
- Vienna University of Technology, Austria

CLEF 2007 Steering Committee

- Maristella Agosti, University of Padua, Italy
- Martin Braschler, Zurich University of Applied Sciences, Switzerland
- Amedeo Cappelli, ISTI-CNR and 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
- Giorgio Di Nunzio, University of Padua, Italy
- David A. Evans, Clairvoyance Corporation, USA
- Marcello Federico, FBK-irst, Trento, Italy
- Nicola Ferro, University of Padua, 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, 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, FBK-irst, Trento, Italy
- Paul McNamee, Johns Hopkins University, USA
- Henning Müller, University and Hospitals of Geneva, Switzerland
- Douglas W. Oard, University of Maryland, USA
- Anselmo Peñas, LSI-UNED, Madrid, Spain
- Maarten de Rijke, University of Amsterdam, Netherlands
- Diana Santos, Linguatca, Sintef, Oslo, Norway
- Jacques Savoy, University of Neuchatel, Switzerland
- Peter Schäuble, 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 2007	1
<i>Carol Peters</i>	

Part I: Multilingual Textual Document Retrieval (Ad Hoc)

CLEF 2007: Ad Hoc Track Overview	13
<i>Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters</i>	

Monolingual

Charles University at CLEF 2007 Ad-Hoc Track	33
<i>Pavel Česka and Pavel Pecina</i>	
Stemming Approaches for East European Languages	37
<i>Ljiljana Dolamic and Jacques Savoy</i>	
Applying Query Expansion Techniques to Ad Hoc Monolingual Tasks with the IR-n System	45
<i>Elisa Noguera and Fernando Llopis</i>	
Bulgarian, Hungarian and Czech Stemming Using YASS	49
<i>Prasenjit Majumder, Mandar Mitra, and Dipasree Pal</i>	
Sampling Precision to Depth 10000 at CLEF 2007	57
<i>Stephen Tomlinson</i>	

Cross-Language: European

Disambiguation and Unknown Term Translation in Cross Language Information Retrieval	64
<i>Dong Zhou, Mark Truran, and Tim Brailsford</i>	
Cross-Language Retrieval with Wikipedia	72
<i>Péter Schönhofen, András Benczúr, István Bóró, and Károly Csalogány</i>	

Cross-Language: Non-European

Cross-Lingual Information Retrieval System for Indian Languages	80
<i>Jagadeesh Jagarlamudi and A. Kumaran</i>	

Bengali, Hindi and Telugu to English Ad-Hoc Bilingual Task at CLEF 2007	88
<i>Sivaji Bandyopadhyay, Tapabrata Mondal, Sudip Kumar Naskar, Asif Ekbal, Rejwanul Haque, and Srinivasa Rao Godhavarthi</i>	
Bengali and Hindi to English CLIR Evaluation	95
<i>Debasis Mandal, Mayank Gupta, Sandipan Dandapat, Pratyush Banerjee, and Sudeshna Sarkar</i>	
Improving Recall for Hindi, Telugu, Oromo to English CLIR	103
<i>Prasad Pingali, Kula Kekeba Tune, and Vasudeva Varma</i>	
Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation.....	111
<i>Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya</i>	
Amharic-English Information Retrieval with Pseudo Relevance Feedback	119
<i>Atelach Alemu Argaw</i>	
Indonesian-English Transitive Translation for Cross-Language Information Retrieval.....	127
<i>Mirna Adriani, Herika Hayurani, and Syandra Sari</i>	

Robust

Robust Retrieval Experiments at the University of Hildesheim	134
<i>Ben Heuwing and Thomas Mandl</i>	
SINAI at CLEF Ad-Hoc Robust Track 2007: Applying Google Search Engine for Robust Cross-Lingual Retrieval	137
<i>F. Martínez-Santiago, A. Montejo-Ráez, and M.Á. García-Cumbreras</i>	
Improving Robustness Using Query Expansion.....	143
<i>Angel F. Zazo, José L. Alonso Berrocal, and Carlos G. Figuerola</i>	
English-to-French CLIR: A Knowledge-Light Approach through Character <i>N</i> -Grams Alignment	148
<i>Jesús Vilares, Michael P. Oakes, and Manuel Vilares</i>	
MIRACLE Progress in Monolingual Information Retrieval at Ad-Hoc CLEF 2007	156
<i>José-Carlos González-Cristóbal, José Miguel Goñi-Menoyo, Julio Villena-Román, and Sara Lana-Serrano</i>	

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

The Domain-Specific Track at CLEF 2007	160
<i>Vivien Petras, Stefan Baerisch, and Maximillian Stempfhuber</i>	

The XTRIEVAL Framework at CLEF 2007: Domain-Specific Track	174
<i>Jens Kürsten, Thomas Wilhelm, and Maximilian Eibl</i>	
Query Translation through Dictionary Adaptation	182
<i>Stephane Clinchant and Jean-Michel Renders</i>	
Experiments in Classification Clustering and Thesaurus Expansion for Domain Specific Cross-Language Retrieval	188
<i>Ray R. Larson</i>	
Domain-Specific IR for German, English and Russian Languages	196
<i>Claire Fautsch, Ljiljana Dolamic, Samir Abdou, and Jacques Savoy</i>	

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

Overview of the CLEF 2007 Multilingual Question Answering Track	200
<i>Danilo Giampiccolo, Pamela Forner, Jesús Herrera, Anselmo Peñas, Christelle Ayache, Corina Forascu, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe</i>	
Overview of the Answer Validation Exercise 2007	237
<i>Anselmo Peñas, Álvaro Rodrigo, and Felisa Verdejo</i>	
Overview of QAST 2007	249
<i>Jordi Turmo, Pere R. Comas, Christelle Ayache, Djamel Mostefa, Sophie Rosset, and Lori Lamel</i>	

Main Task: Mono- and Bilingual QA

Question Answering with Joost at CLEF 2007	257
<i>Gosse Bouma, Geert Kloosterman, Jori Mur, Gertjan van Noord, Lonneke van der Plas, and Jörg Tiedemann</i>	
What Happened to Esfinge in 2007?	261
<i>Luís Miguel Cabral, Luís Fernando Costa, and Diana Santos</i>	
Coreference Resolution for Questions and Answer Merging by Validation	269
<i>Sven Hartrumpf, Ingo Glöckner, and Johannes Leveling</i>	
Multilingual Question Answering through Intermediate Translation: LCC's PowerAnswer at QA@CLEF 2007	273
<i>Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Thomas d'Silva, and Dan Moldovan</i>	
RACAI's Question Answering System at QA@CLEF2007	284
<i>Dan Tufiş, Dan Ştefănescu, Radu Ion, and Alexandru Ceaşu</i>	

DFKI-LT at QA@CLEF 2007	292
<i>Bogdan Sacaleanu, Günter Neumann, and Christian Spurk</i>	
University of Wolverhampton at CLEF 2007.....	300
<i>Georgiana Puşcaşu and Constantin Orăsan</i>	
Bilingual Question Answering Using CINDI-QA at QA@CLEF 2007....	308
<i>Chedid Haddad and Bipin C. Desai</i>	
The University of Évora's Participation in QA@CLEF-2007	316
<i>José Saias and Paulo Quaresma</i>	
Web-Based Anaphora Resolution for the QUASAR Question Answering System	324
<i>Davide Buscaldi, Yassine Benajiba, Paolo Rosso, and Emilio Sanchis</i>	
A Lexical Approach for Spanish Question Answering	328
<i>Alberto Téllez, Antonio Juárez, Gustavo Hernández, Claudia Denicia, Esaú Villatoro, Manuel Montes, and Luis Villaseñor</i>	
Finding Answers Using Resources in the Internet	332
<i>Mirna Adriani and Septian Adiwibowo</i>	
UAIC Romanian QA System for QA@CLEF	336
<i>Adrian Iftene, Diana Trandabăt, Ionuţ Pistol, Alex Moruz, Alexandra Balahur, Diana Cotelea, Iustin Dornescu, Iuliana Drăghici, and Dan Cristea</i>	
The University of Amsterdam's Question Answering System at QA@CLEF 2007	344
<i>Valentin Jijkoun, Katja Hofmann, David Ahn, Mahboob Alam Khalid, Joris van Rantwijk, Maarten de Rijke, and Erik Tjong Kim Sang</i>	
Combining Wikipedia and Newswire Texts for Question Answering in Spanish	352
<i>César de Pablo-Sánchez, José L. Martínez-Fernández, Ana González-Ledesma, Doaa Samy, Paloma Martínez, Antonio Moreno-Sandoval, and Harith Al-Jumaily</i>	
QA@L ² F, First Steps at QA@CLEF.....	356
<i>Ana Mendes, Luísa Coheur, Nuno J. Mamede, Ricardo Ribeiro, Fernando Batista, and David Martins de Matos</i>	
Priberam's Question Answering System in QA@CLEF 2007	364
<i>Carlos Amaral, Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, Cláudia Pinto, and Daniel Vidal</i>	

Answer Validation Exercise (AVE)

Combining Logic and Aggregation for Answer Selection	372
<i>Ingo Glöckner</i>	
On the Application of Lexical-Syntactic Knowledge to the Answer Validation Exercise	377
<i>Óscar Ferrández, Daniel Micol, Rafael Muñoz, and Manuel Palomar</i>	
Combining Lexical Information with Machine Learning for Answer Validation at QA@CLEF 2007	381
<i>M.Á. García-Cumbreras, J.M. Perea-Ortega, F. Martínez-Santiago, and L. Alfonso Ureña-López</i>	
Using Recognizing Textual Entailment as a Core Engine for Answer Validation	387
<i>Rui Wang and Günter Neumann</i>	
A Supervised Learning Approach to Spanish Answer Validation	391
<i>Alberto Téllez-Valero, Manuel Montes-y-Gómez, and Luis Villaseñor-Pineda</i>	
UAIC Participation at AVE 2007	395
<i>Adrian Iftene and Alexandra Balahur-Dobrescu</i>	
UNED at Answer Validation Exercise 2007	404
<i>Álvaro Rodrigo, Anselmo Peñas, and Felisa Verdejo</i>	

Question Answering on Speech Transcription (QAST)

Adapting QA Components to Mine Answers in Speech Transcripts	410
<i>Günter Neumann and Rui Wang</i>	
The LIMSI Participation in the QAST Track	414
<i>Sophie Rosset, Olivier Galibert, Gilles Adda, and Eric Bilinski</i>	
Robust Question Answering for Speech Transcripts Using Minimal Syntactic Analysis	424
<i>Pere R. Comas, Jordi Turmo, and Mihai Surdeanu</i>	

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

Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task . . .	433
<i>Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller</i>	

Overview of the ImageCLEF 2007 Object Retrieval Task	445
<i>Thomas Deselaers, Allan Hanbury, Ville Viitaniemi, András Benczúr, Mátyás Brendel, Bálint Daróczy, Hugo Jair Escalante Balderas, Theo Gevers, Carlos Arturo Hernández Gracidas, Steven C.H. Hoi, Jorma Laaksonen, Mingjing Li, Heidy Marisol Marín Castro, Hermann Ney, Xiaoguang Rui, Nicu Sebe, Julian Stöttinger, and Lei Wu</i>	
Overview of the ImageCLEFmed 2007 Medical Retrieval and Medical Annotation Tasks	472
<i>Henning Müller, Thomas Deselaers, Thomas M. Deserno, Jayashree Kalpathy-Cramer, Eugene Kim, and William Hersh</i>	
ImageCLEFphoto	
FIRE in ImageCLEF 2007: Support Vector Machines and Logistic Models to Fuse Image Descriptors for Photo Retrieval	492
<i>Tobias Gass, Tobias Weyand, Thomas Deselaers, and Hermann Ney</i>	
MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval	500
<i>Julio Villena-Román, Sara Lana-Serrano, José Luis Martínez-Fernández, and José Carlos González-Cristóbal</i>	
Using an Image-Text Parallel Corpus and the Web for Query Expansion in Cross-Language Image Retrieval	504
<i>Yih-Chen Chang and Hsin-Hsi Chen</i>	
SINAI System: Combining IR Systems at ImageCLEFPhoto 2007	512
<i>M.Á. García-Cumbreras, M.C. Díaz-Galiano, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña-López</i>	
Multimodal Retrieval by Text–Segment Biclustering	518
<i>András Benczúr, István Bíró, Mátyás Brendel, Károly Csalogány, Bálint Daróczy, and Dávid Siklósi</i>	
Analysing an Approach to Information Retrieval of Visual Descriptions with IR-n, a System Based on Passages	522
<i>Sergio Navarro, Fernando Llopis, Rafael Muñoz Guillena, and Elisa Noguera</i>	
DCU and UTA at ImageCLEFPhoto 2007	530
<i>Anni Järvelin, Peter Wilkins, Tomasz Adamek, Eija Airio, Gareth J.F. Jones, Alan F. Smeaton, and Eero Sormunen</i>	

Cross-Language and Cross-Media Image Retrieval: An Empirical Study at ImageCLEF2007	538
<i>Steven C.H. Hoi</i>	
Towards Annotation-Based Query and Document Expansion for Image Retrieval	546
<i>Hugo Jair Escalante, Carlos Hernández, Aurelio López, Heidy Marín, Manuel Montes, Eduardo Morales, Enrique Sucar, and Luis Villaseñor</i>	
Content-Based Image Retrieval Using Combined 2D Attribute Pattern Spectra	554
<i>Florence Tushabe and Michael. H.F. Wilkinson</i>	
Text-Based Clustering of the ImageCLEFphoto Collection for Augmenting the Retrieved Results	562
<i>Osama El Demerdash, Leila Kosseim, and Sabine Bergler</i>	
Trans-Media Pseudo-Relevance Feedback Methods in Multimedia Retrieval	569
<i>Stephane Clinchant, Jean-Michel Renders, and Gabriela Csurka</i>	
ImageCLEFmed	
Cue Integration for Medical Image Annotation	577
<i>Tatiana Tommasi, Francesco Orabona, and Barbara Caputo</i>	
Multiplying Concept Sources for Graph Modeling	585
<i>Loic Maisonnasse, Eric Gaussier, and Jean Pierre Chevallet</i>	
MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval	593
<i>Julio Villena-Román, Sara Lana-Serrano, and José Carlos González-Cristóbal</i>	
MIRACLE at ImageCLEFanot 2007: Machine Learning Experiments on Medical Image Annotation	597
<i>Sara Lana-Serrano, Julio Villena-Román, José Carlos González-Cristóbal, and José Miguel Goñi-Menoyo</i>	
Integrating MeSH Ontology to Improve Medical Information Retrieval	601
<i>M.C. Díaz-Galiano, M.Á. García-Cumbreras, M.T. Martín-Valdivia, A. Montejo-Ráez, and L.A. Ureña-López</i>	
Speeding Up IDM without Degradation of Retrieval Quality	607
<i>Michael Springmann and Heiko Schuldt</i>	

Content-Based Medical Image Retrieval Using Low-Level Visual Features and Modality Identification	615
<i>Juan C. Caicedo, Fabio A. Gonzalez, and Eduardo Romero</i>	
Medical Image Retrieval and Automatic Annotation: OHSU at ImageCLEF 2007	623
<i>Jayashree Kalpathy-Cramer and William Hersh</i>	
Using Bayesian Network for Conceptual Indexing: Application to Medical Document Indexing with UMLS Metathesaurus	631
<i>Thi Hoang Diem Le, Jean-Pierre Chevallet, and Joo Hwee Lim</i>	
Baseline Results for the ImageCLEF 2007 Medical Automatic Annotation Task Using Global Image Features	637
<i>Mark O. Güld and Thomas M. Deserno</i>	
Evaluation of Automatically Assigned MeSH Terms for Retrieval of Medical Images	641
<i>Miguel E. Ruiz and Aurélie Névél</i>	

ImageCLEF photo and med

University and Hospitals of Geneva Participating at ImageCLEF 2007	649
<i>Xin Zhou, Julien Gobeill, Patrick Ruch, and Henning Müller</i>	
An Interactive and Dynamic Fusion-Based Image Retrieval Approach by CINDI	657
<i>M.M. Rahman, B.C. Desai, and P. Bhattacharya</i>	
Using Pseudo-Relevance Feedback to Improve Image Retrieval Results	665
<i>Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem</i>	

Part V: Cross-Language Speech Retrieval (CL-SR)

Overview of the CLEF-2007 Cross-Language Speech Retrieval Track	674
<i>Pavel Pecina, Petra Hoffmannová, Gareth J.F. Jones, Ying Zhang, and Douglas W. Oard</i>	
A Dirichlet-Smoothed Bigram Model for Retrieving Spontaneous Speech	687
<i>Matthew Lease and Eugene Charniak</i>	
Model Fusion Experiments for the CLSR Task at CLEF 2007	695
<i>Muath Alzghool and Diana Inkpen</i>	

Dublin City University at CLEF 2007: Cross-Language Speech Retrieval Experiments	703
<i>Ying Zhang, Gareth J.F. Jones, and Ke Zhang</i>	
What Can and Cannot Be Found in Czech Spontaneous Speech Using Document-Oriented IR Methods—UWB at CLEF 2007 CL-SR Track.....	712
<i>Pavel Ircing, Josef Psutka, and Jan Vavruška</i>	
Using Information Gain to Filter Information in CLEF CL-SR Track ...	719
<i>M.C. Díaz-Galiano, M.T. Martín-Valdivia, M.Á. García-Cumbreras, and L.A. Ureña-López</i>	

Part VI: Multilingual Web Retrieval (WebCLEF)

Overview of WebCLEF 2007	725
<i>Valentin Jijkoun and Maarten de Rijke</i>	
Segmentation of Web Documents and Retrieval of Useful Passages	732
<i>Carlos G. Figuerola, José L. Alonso Berrocal, and Angel F. Zazo Rodríguez</i>	
Using Centrality to Rank Web Snippets.....	737
<i>Valentin Jijkoun and Maarten de Rijke</i>	
Using Web-Content for Retrieving Snippets	742
<i>Okky Hendriansyah, Tri Firgantoro, and Mirna Adriani</i>	

Part VII: Cross-Language Geographical Retrieval (GeoCLEF)

GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview	745
<i>Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie</i>	
Inferring Location Names for Geographic Information Retrieval	773
<i>Johannes Leveling and Sven Hartrumpf</i>	
GeoParsing Web Queries	781
<i>Rocio Guillén</i>	
MIRACLE at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information.....	786
<i>Sara Lana-Serrano, Julio Villena-Román, José Carlos González-Cristóbal, and José Miguel Goñi-Menoyo</i>	

Relevance Measures Using Geographic Scopes and Types	794
<i>Geoffrey Andogah and Gosse Bouma</i>	
Using Geographic Signatures as Query and Document Scopes in Geographic IR	802
<i>Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva</i>	
Cheshire at GeoCLEF 2007: Retesting Text Retrieval Baselines	811
<i>Ray R. Larson</i>	
On the Relative Importance of Toponyms in GeoCLEF	815
<i>Davide Buscaldi and Paolo Rosso</i>	
Filtering for Improving the Geographic Information Search	823
<i>José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, and L.A. Ureña-López</i>	
TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier	830
<i>Daniel Ferrés and Horacio Rodríguez</i>	
TALP at GeoQuery 2007: Linguistic and Geographical Analysis for Query Parsing	834
<i>Daniel Ferrés and Horacio Rodríguez</i>	
Applying Geo-feedback to Geographic Information Retrieval	838
<i>Mirna Adriani and Nasikhin</i>	
Exploring LDA-Based Document Model for Geographic Information Retrieval	842
<i>Zhisheng Li, Chong Wang, Xing Xie, Xufa Wang, and Wei-Ying Ma</i>	
Mono- and Crosslingual Retrieval Experiments with Spatial Restrictions at GeoCLEF 2007	850
<i>Ralph Kösle, Ben Hewwing, Thomas Mandl, and Christa Womser-Hacker</i>	
GIR Experiments with Forostar	856
<i>Simon Overell, João Magalhães, and Stefan Rüger</i>	

Part VIII: CLEF in Other Evaluations

CLEF at MorphoChallenge

Morpho Challenge Evaluation Using a Linguistic Gold Standard	864
<i>Mikko Kurimo, Mathias Creutz, and Matti Varjokallio</i>	
Simple Morpheme Labelling in Unsupervised Morpheme Analysis	873
<i>Delphine Bernhard</i>	

Unsupervised and Knowledge-Free Morpheme Segmentation and Analysis	881
<i>Stefan Bordag</i>	
Unsupervised Acquiring of Morphological Paradigms from Tokenized Text	892
<i>Daniel Zeman</i>	
ParaMor: Finding Paradigms across Morphology	900
<i>Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin</i>	
 CLEF at SemEval 2007	
SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval	908
<i>Eneko Agirre, Oier Lopez de Lacalle, Bernardo Magnini, Arantra Otegi, German Rigau, and Piek Vossen</i>	
Author Index	919

What Happened in CLEF 2007

Carol Peters

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

Abstract. The organization of the CLEF 2007 evaluation campaign is described and details are provided concerning the tracks, test collections, evaluation infrastructure, and participation. The main results are commented and future evolutions in the organization of CLEF are discussed.

1 Introduction

The objective of the Cross Language Evaluation Forum is to promote research in the field of multilingual system development. This is done through the organization 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.

This volume contains a series of papers describing the research activities and experiments that were conducted under the umbrella of the CLEF 2007 campaign. The main features of this campaign are briefly outlined below in order to provide the necessary background to these papers. In the final sections, we comment on the main results obtained and discuss the changes that will be made in the organization of CLEF in the next two years.

2 Tracks and Tasks in CLEF 2007

CLEF 2007 offered seven 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 retrieval on structured scientific data (Domain-Specific)
- multiple language question answering (QA@CLEF)
- cross-language retrieval in image collections (ImageCLEF)
- cross-language speech retrieval (CL-SR)

- multilingual retrieval of Web documents (WebCLEF)
- cross-language geographical retrieval (GeoCLEF).

These tracks are mainly the same as those offered in CLEF 2006 with the exclusion of an interactive track¹, however many of the tasks offered are new. In addition to the tracks organized within CLEF 2007, two external evaluation activities (SemEval 2007 and Morpho Challenge 2007) included tasks which adopted CLEF resources and test collections. The results of these tasks are also reported in this volume.

Cross-Language Text Retrieval (Ad Hoc): This year, this track offered mono- and bilingual tasks on target collections for central European languages (Bulgarian, Czech² and Hungarian). Similarly to last year, a bilingual task encouraging system testing with non-European languages against English documents was offered. Topics were made available in Amharic, Chinese, Oromo and Indonesian. A special sub-task regarding Indian language search against an English target collection was also organized with the assistance of a number of Indian research institutes, responsible for the preparation of the topics. The languages offered were Hindi, Bengali, Tamil, Telugu and Marathi. A "robust" task was again offered, emphasizing the importance of reaching a minimal performance for all topics instead of high average performance. Robustness is a key issue for the transfer of CLEF research into applications. The 2007 robust task involved three languages often used in previous CLEF campaigns (English, French, Portuguese). The track was coordinated jointly by ISTI-CNR, Pisa, Italy, and the Universities of Padua, Italy, and Hildesheim, Germany.

Cross-Language Scientific Data Retrieval (Domain-Specific): Mono- and cross-language domain-specific retrieval was studied in the domain of social sciences using structured data (e.g. bibliographic data, keywords, and abstracts) from scientific reference databases. The target collections provided were in English, German and Russian and topics were offered in the same languages. This track was coordinated by InformationsZentrum Sozialwissenschaften, Bonn, Germany.

Multilingual Question Answering (QA@CLEF): This track proposed both main and pilot tasks. The main task scenario was topic-related QA, where the questions are grouped by topics and may contain anaphoric references one to the others. The answers were retrieved from heterogeneous document collections, i.e. news articles and Wikipedia. Many sub-tasks were set up, monolingual – where the questions and the target collections searched for answers are in the same language – and bilingual – where source and target languages are different. Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish were offered as target languages; query languages used in the bilingual tasks depended on demand (see the track overview for details). Following the positive response at QA@CLEF 2006, the Answer Validation Exercise (AVE) was proposed again. A new pilot task was also offered: Question

¹ From CLEF 2001 through CLEF 2006, we have offered an interactive track. Unfortunately, this year, the track was suspended due to other commitments of the organisers. Owing to the importance of user intervention in cross-language IR, we intend to re-propose and strengthen the interactive activity in CLEF 2008.

² New this year.

Answering on Speech Transcript (QAST), in which the answers to factual questions have to be extracted from spontaneous speech transcriptions (manual and automatic transcriptions) coming from different human interaction scenarios. The track was organized by several institutions (one for each source language) and jointly coordinated by CELCT, Trento, Italy, Universidad Nacional de Educación a Distancia, Madrid, Spain, and Universitat Politècnico de Catalunya, Barcelona, Spain.

Cross-Language Retrieval in Image Collections (ImageCLEF): This track evaluated retrieval of images described by text captions in several languages; both text and image retrieval techniques were exploitable. Four challenging tasks were offered: (i) multilingual ad-hoc retrieval (collection with mixed English/German/Spanish annotations, queries in more languages), (ii) medical image retrieval (case notes in English/French/German; visual, mixed, semantic queries in same languages), (iii) hierarchical automatic image annotation for medical images (fully categorized in English and German, purely visual task), (iv) photographic annotation through detection of objects in images (using the same collection as (i) with a restricted number of objects, a purely visual task). Image retrieval was not required for all tasks and a default visual and textual retrieval system was made available for participants. The track coordinators were Sheffield University, UK, and the University and University Hospitals of Geneva, Switzerland. Oregon Health and Science University, USA, Victoria University, Melbourne, Australia, RWTH Aachen, Germany and Vienna University of Technology, Austria, collaborated in the task organization.

Cross-Language Speech Retrieval (CL-SR): The focus was on searching spontaneous speech from oral history interviews rather than news broadcasts. The test collection created for the track is a subset of a large archive of videotaped oral histories from survivors, liberators, rescuers and witnesses of the Holocaust created by the Survivors of the Shoah Visual History Foundation. Automatic Speech Recognition (ASR) transcripts and both automatically and manually assigned thesaurus terms were available as part of the collection. In 2006 the CL-SR track included search collections of conversational English and Czech speech using six languages (Czech, Dutch, English, French, German and Spanish). In CLEF 2007 additional topics were added for the Czech speech collection. The track was coordinated by University of Maryland, USA, Dublin City University, Ireland, and Charles University, Czech Republic.

Multilingual Web Retrieval (WebCLEF): The WebCLEF 2007 task combined insights gained from previous editions of WebCLEF 2005–2006 and the WiQA 2006 pilot, and went beyond the navigational queries considered at WebCLEF 2005 and 2006. At WebCLEF 2007 so-called undirected informational search goals were considered in a web setting: “I want to learn anything/everything about my topic.” The track was coordinated by University of Amsterdam, The Netherlands.

Cross-Language Geographical Retrieval (GeoCLEF): The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval for topics with a geographic specification. GeoCLEF 2007 consisted of two sub tasks. A search task ran for the third time and a query classification task was organized for the first. The

document collections were in English, German and Portuguese and topics were prepared in these languages plus Spanish. For the classification task, a query log from a search engine was provided and the groups needed to identify the queries with a geographic scope and the geographic components within the local queries. The track was coordinated jointly by UC Berkeley, USA, the Universities of Sheffield, UK, and Hildesheim, Germany, Linguatca SINTEF, Norway and Microsoft Asia, China.

More complete details on the technical infrastructure and the organization of these tracks can be found in the track overview reports in this volume, collocated at the beginning of the relevant sections.

3 Test Collections

Test collections consist of collections of documents together with sets of topics or queries designed to evaluate the participating systems for particular performance aspects according to the focus of the specific task, and sets of relevance judgments created in order to be able to assess and analyse the results.

A number of different document collections were used in CLEF 2007 to build the test collections:

- CLEF multilingual comparable corpus of more than 3 million news documents in 13 languages; new data was added this year for Czech, Bulgarian and English (see Table 1); Parts of this collections were used in the Ad-Hoc, QuestionAnswering, and GeoCLEF tracks.
- The GIRT-4 social science database in English and German (over 300,000 documents), Cambridge Sociological Abstracts in English (20,000 documents) and the Russian ISSS collection for sociology and economics (approx. 150,000 docs). These collections were used in the domain-specific track.
- The ImageCLEF track used collections for both general photographic and medical image retrieval:
 - IAPR TC-12 photo database of 20,000 colour photographs with captions in English, German and Spanish; PASCAL VOC 2006 training data (new this year);
 - ImageCLEFmed radiological database consisting of 6 distinct datasets – 2 more than last year; IRMA collection in English and German of 12,000 classified images for automatic medical image annotation
- Malach collection of spontaneous conversational speech derived from the Shoah archives in English (more than 750 hours) and Czech (approx 500 hours). This collection was used in the speech retrieval track.
- EuroGOV, a multilingual collection of about 3.5M webpages, containing documents in many languages crawled from European governmental sites, used in the WebCLEF track.

The coordinators of each track and/or task were responsible for creating the sets of topics in many different languages, depending on the task, and for preparing the relevance judgments according to previously decided criteria.

Table 1. Sources and dimensions of the CLEF 2007 multilingual comparable corpus, new collections indicated in bold

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
Bulgarian: Novinar 2002	2007	48	18,086	NA
Czech: Mladna frontaDnes 2002	2007	143	68,842	NA
Czech: Lidove Noviny 2002	2007	35	12,893	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: LA Times 2002	2007	434	135,153	NA
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).

4 Technical Infrastructure

The DIRECT³ system, designed and developed at the University of Padua manages the CLEF test data plus results submission and analyses for the ad hoc, question answering and geographic IR tracks.

DIRECT also supports the production, maintenance, enrichment and interpretation of scientific data for subsequent in-depth evaluation studies. It is thus responsible for:

- track set-up, harvesting of documents, management of registrations;
- 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.

• Brown Univ., USA	• Johns Hopkins Univ., USA	• Univ. Essex, Portugal	•
• California State Univ. San Marcos, USA	• Language Computer Corp., USA	• Univ. Freiburg, Germany	•
• Charles Univ., Prague, Czech Rep.	• LIMSI-CNRS, France	• Univ. de Hacques Girona, CH	•
• Dicosmis & Madrid Univ. Consortium, Spain	• Linguistics Sinfel, Norway	• Univ. Groningen Inf Sci, The Netherlands	•
• Chung Yuen Univ., Taiwan	• Linguit Ltd, UK	• Univ. Hagen - IICS, Germany	•
• DFKI-Artificial Intelligence, Germany	• Microsoft Asia	• Univ. Hildesheim, Germany	•
• Dokuz Eylul Univ., Turkey	• Microsoft India	• Univ. Indonesia, Indonesia	•
• Dublin City Univ., Ireland	• MIRM - LBG, Grenoble, France	• Univ. Jaen, Spain	•
• Fondazione Bruno Kessler, Italy	• Nat. Inst. Informatics, Japan	• Univ. Lege, Belgium	•
• Helsinki Univ. of Technology, Finland	• Nat. Taiwan Univ.	• Univ. Lisbon, Portugal	•
• Hungarian Acad. Sci.	• Open Text Corp., Canada	• Univ. Macquarie, Australia	•
• IDIAP Research Inst., Switzerland	• Oregon Health & Sci Univ., USA	• Univ. National Colombia	•
• Imperial College, London, UK	• Paborca Informatica, Portugal	• UNFuchatel, Switzerland	•
• In Mar Autrofinica, Optica, Electronica, Mexico	• Research Inst. for AI of Romanian Academy	• Univ. Nottingham, U.K.	•
• Indian Statistical Inst., India	• RWTH Aachen CS., Germany	• Univ. Ottawa - IT & Eng., Canada	•
• Indian Inst. Technology Bombay	• RWTH Aachen - Med Inf., Germany	• Univ. Politecnica Catalunya, Spain	•
• Indian Inst. Technology Kharagpur	• SUNY Buffalo, USA	• Univ. Politecnica Valencia, Spain	•
• Inst. Informatica Research, Singapore	• SYNAPSE Development, France	• Univ. Porto, Portugal	•
• Inst. Superior Tecnico, Portugal	• Tech Univ. Chemnitz, Germany	• Univ. Salamanca, Spain	•
• ITAL-CNRS, Singapore	• Tokyo Inst. Technology, Japan	• Univ. Stockholm, Sweden	•
• IRIT-SIG Toulouse, France	• Univ. Alcala, Spain	• Univ. Tampere, Finland	•
• Jyväskylä University, Kallio, India	• Univ. AII Cuza Iasi, Romania	• Univ. Wolverhampton, UK	•
	• Univ. Amsterdam, NL	• UC Berkeley, USA	•
	• Univ. Basel, Switzerland	• UNED LSI, Spain	•
	• Univ. Chicago, USA	• Univ. West Bohemia, Czech Rep.	•
	• Univ. Concordia - CINDI, Canada	• Vietnam Univ. Technology, Austria	•
	• Univ. Concordia - CLAX, Canada	• Xerox XRCCE, France	•

Fig. 1. CLEF 2007 Participation

5 Participation

A total of 81 groups submitted runs in CLEF 2007, slightly down from the 90 groups of CLEF 2006: 51(59.5) from Europe, 14(14.5) from North America; 14(10) from Asia,

³ <http://direct.dei.unipd.it/>

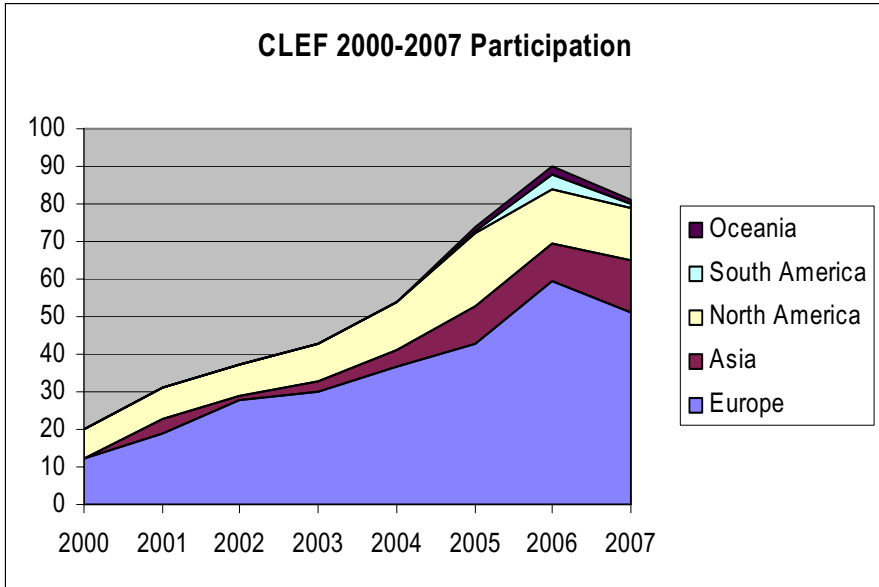


Fig. 2. CLEF 2000 – 2007: Variation in Participation

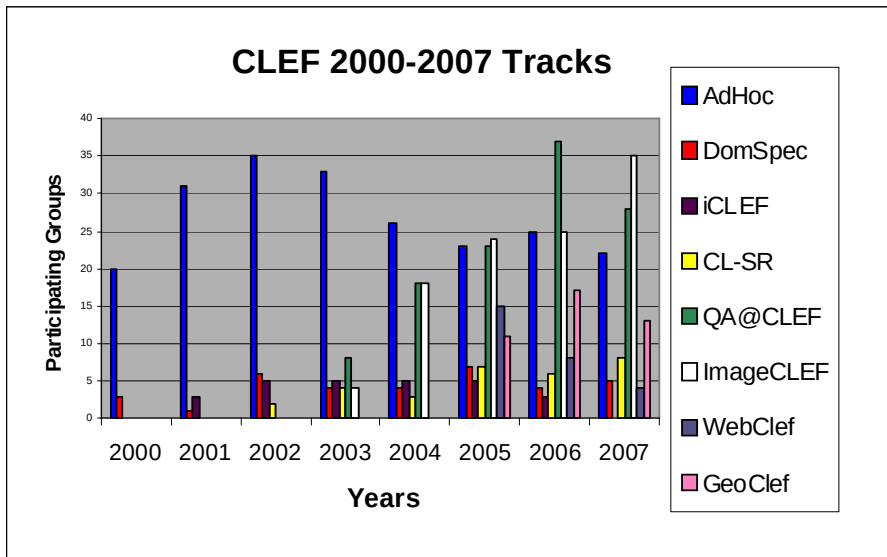


Fig. 3. CLEF 2000 – 2007: Participation per Track in Tracks

1(4) from South America and 1(1) from Australia. The breakdown of participation of groups per track is as follows: Ad Hoc 22(25); Domain-Specific 5(4); QAatCLEF 28(37); ImageCLEF 35(25); CL-SR 8(6); WebCLEF 4(8); GeoCLEF 13(17)⁴. Fig. 1 lists the participants, Fig. 2 shows variations in participation over the years and Fig. 3 shows the shift in focus as new tracks have been added.

These figures show that, while there is a constant increase in interest in the ImageCLEF track, there is a decrease in popularity of the question answering and web tracks. Although the fluctuation in QA does not appear of great significance, the apparent lack of interest in WebCLEF is surprising. With the importance of Internet and web search engines, a greater participation in this task is to be expected. The large numbers for ImageCLEF also give rise to some discussion. The defining feature of CLEF is its multilinguality; ImageCLEF is perhaps the least multilingual of the CLEF tracks as much of the work is done in a language-independent context.

It should be noted that this volume also includes reports from two separate evaluation initiatives which used CLEF data for certain tasks – thus in 2007 the impact of CLEF spread beyond the strict boundaries of the CLEF evaluation campaigns.

6 Main Results

The main results of the CLEF activity over the years can be summarized in the following points:

- Stimulation of research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval
- Study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- Documented improvement in system performance for cross-language text retrieval systems
- Creation of a large set of empirical data about multilingual information access from the user perspective
- Quantitative and qualitative evidence with respect to best practice in cross-language system development
- Creation of important, reusable test collections for system benchmarking
- Building of a strong, multidisciplinary research community.

These results are confirmed in CLEF 2007.

7 From CLEF to TrebleCLEF

So far, CLEF has been a forum where researchers can perform experiments, discuss results and exchange ideas; most of the results have been published but the extensive CLEF-related literature is mainly intended for the academic community. Contacts with interested application communities have been notably lacking.

In fact, evaluation campaigns have their limitations. They tend to focus on aspects of system performance that can be measured easily in an objective setting (e.g. precision and recall) and to ignore others that are equally important for overall system development. Thus, while in CLEF, much attention has been paid to improving performance in terms of

⁴ Last year's figures are between brackets.

the ranking of results through the refining of query expansion procedures, term weighting schemes, algorithms for the merging of results, equally important criteria of speed, stability, usability have been mainly ignored. Also for any MLIA system, the results must be presented in an understandable and useful fashion. The user interface implementation needs to be studied very carefully according to the particular user profile. Such aspects tend to be neglected in traditional evaluation campaigns.

We have thus decided to launch a new activity which aims at building on and extending the results already achieved by CLEF. This activity, called TrebleCLEF⁵, will stimulate the development of operational MLIA systems rather than research prototypes. TrebleCLEF intends to promote research, development, implementation and industrial take-up of multilingual, multimodal information access functionality in the following ways:

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to stimulate R&D to meet the requirements of the user and application communities, with particular focus on the following key areas:
 - user modeling, e.g. what are the requirements of different classes of users when querying multilingual information sources;
 - language-specific experimentation, e.g. looking at differences across languages in order to derive best practices for each language, best practices for the development of system components, and best practices for MLIA systems as a whole;
 - results presentation, e.g. how can results be presented in the most useful and comprehensible way to the user.
- by constituting a scientific forum for the MLIA community of researchers enabling them to meet and discuss results, emerging trends, new directions:
 - providing a scientific digital library with tools for analyzing, comparing, and citing the scientific data of an evaluation campaign, as well as curating, preserving, annotating, enriching, and promoting the re-use of them;
- by acting as a virtual centre of competence providing a central reference point for anyone interested in studying or implementing MLIA functionality and encouraging the dissemination of information:
 - making publicly available sets of guidelines on best practices in MLIA (e.g. what stemmer to use, what stop list, what translation resources, how best to evaluate, etc., depending on the application requirements);
 - making tools and resources used in the evaluation campaigns freely available to a wider public whenever possible; otherwise providing links to where they can be acquired;
 - organizing workshops, and/or tutorials and training sessions.

To sum up, TrebleCLEF will not only sponsor R&D and evaluation in the multilingual retrieval context but will focus on those aspects of system implementation that have been somewhat neglected so far with the aim of preparing an exhaustive set of best practice recommendations addressing the issues involved from both the system

⁵ TrebleCLEF is a 7FP Coordination Action under the IST programme; it began activity in January 2008. The Consortium is composed of five academic partners and two important centres: ISTI-CNR, Italy; University of Padua, Italy, University of Sheffield, UK; Zurich University of Applied Sciences, Switzerland; UNED, Spain; CELCT, Italy, ELDA, France. See www.trebleclef.eu.

and the user perspective. The goal is to disseminate the research findings to system developers encouraging easy take up of MLIA technology by the application communities.

8 CLEF 2008

At the time of writing the CLEF 2008 campaign is in full swing. There are seven main tracks plus two new pilot tasks for cross-language video retrieval (VideoCLEF) and multilingual information filtering (INFILE@CLEF)⁶. Although most of the main tracks are the same as last year (with the exception of the speech retrieval track which is not offered in 2008), the majority of tracks include new tasks. Several new collections have been added and, in keeping with the desire to move closer to the application communities, the Ad Hoc track has organized a task in collaboration with The European Library aimed at testing monolingual and cross-language retrieval on library catalog archives. CLEF 2008 also provides testing on Persian (in Ad Hoc) and Arabic (INFILE) collections. This is the first time that CLEF has provided document collections in languages from outside Europe. Finally, particular attention is being given to the interactive track which is focused on interactive experiments in a multilingual context using the Flickr collections and search log analysis studying correlations between search success and search strategies, or language skills, etc..

The results will be presented and discussed at the CLEF 2008 workshop, to be held in Aarhus, Denmark, 17 – 19 September 2008.

Acknowledgements

It would be impossible to run the CLEF evaluation initiative and organize the annual workshops without considerable assistance from many groups. 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 have been involved in the coordination of the 2007 campaigns. A list of the main institutions involved is given at the beginning of this volume. Here below, let me thank the people mainly responsible for the coordination of the different tracks:

- Giorgio Di Nunzio, Nicola Ferro and Thomas Mandl for the Ad Hoc Track
- Vivien Petras, Stefan Baerisch, Maximillian Stempfhuber for the Domain-Specific track
- Bernardo Magnini, Danilo Giampiccolo Christelle Ayache, Petya Osenova, Anselmo Peñas, Maarten de Rijke, Bogdan Sacaleanu, Diana Santos and Richard Sutcliffe for QA@CLEF
- Allan Hanbury, Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Jayashree Kalpathy-Cramer, and William Hersch for ImageCLEF
- Douglas W. Oard, Gareth J. F. Jones, and Pavel Pecina for CL-SR
- Valentin Jijkoun and Maarten de Rijke for Web-CLEF

⁶ See www.clef-campaign.org/2008/2008_agenda.html

- Thomas Mandl, Fredric C. Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie for GeoCLEF.

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 for their work on Chinese.

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, Novinar OD Sofia, and the BulTreeBank Project, Linguistic Modelling Laboratory, IPP, Bulgarian Acad. Sci, for the Bulgarian newspaper documents
- Mafra a.s. and Lidové Noviny a.s. for the Czech newspaper data
- 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
- The Radiology Dept. of the University Hospitals of Geneva for the Casimage database and the PEIR (Pathology Education Image Resource) for the images and the HEAL (Health Education Assets Library) for the Annotation of the Peir dataset.

- Aachen University of Technology (RWTH), Germany for the IRMA database of annotated medical images
- Mallinkrodt Institute of Radiology for permission to use their nuclear medicine teaching file
- University of Basel's Pathologic project for their Pathology teaching file
- Michael Grubinger, administrator of the IAPR Image Benchmark, Clement Leung who initiated and supervised the IAPR Image Benchmark Project, and André Kiwitz, the Managing Director of Viventura for granting access to the image database and the raw image annotations of the tour guides.
- 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 and Valeria Quochi for their assistance in the organization of the CLEF 2007 Workshop.

CLEF 2007: 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, Pisa, Italy
carol.peters@isti.cnr.it

Abstract. We describe the objectives and organization of the CLEF 2007 Ad Hoc track and discuss the main characteristics of the tasks offered to test monolingual and cross-language textual document retrieval systems. The track was divided into two streams. The main stream offered mono- and bilingual tasks on target collections for central European languages (Bulgarian, Czech and Hungarian). Similarly to last year, a bilingual task that encouraged system testing with non-European languages against English documents was also offered; this year, particular attention was given to Indian languages. The second stream, designed for more experienced participants, offered mono- and bilingual “robust” tasks with the objective of privileging experiments which achieve good stable performance over all queries rather than high average performance. These experiments re-used CLEF test collections from previous years in three languages (English, French, and Portuguese). The performance achieved for each task is presented and discussed.

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. Similarly to last year, the CLEF 2007 ad hoc track was structured in two streams. The main stream offered mono- and bilingual retrieval tasks on target collections for central European languages plus a bilingual task encouraging system testing with non-European languages against English documents. The second stream, designed for more experienced participants, was the “robust task”, aimed at finding documents for very difficult queries. It used test collections developed in previous years.

The **Monolingual** and **Bilingual** tasks were principally offered for Bulgarian, Czech and Hungarian target collections. Additionally, a bilingual task was offered to test querying with non-European language queries against an English target collection. As a result of requests from a number of Indian research institutes, a special sub-task for Indian languages was offered with topics in Bengali, Hindi,

Marathi, Tamil and Telugu. 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 proposed mono- and bilingual experiments using the test collections built over the last six CLEF campaigns. Collections and topics in English, Portuguese and French were used. The goal of the robust analysis is to improve the user experience with a retrieval system. Poor performing topics are more serious for the user than performance losses in the middle and upper interval. The robust task gives preference to systems which achieve a minimal level for all topics. The measure used to ensure this is the geometric mean over all topics. The robust task intends to evaluate stable performance over all topics instead of high average performance. Experiments are offered with a larger topic set.

This was the first year since CLEF began that we have not offered a **Multilingual** ad hoc task (ie searching a target collection in multiple languages).

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 this year. The main stream used national newspaper documents from 2002 as the target collections, creating sets of new topics and making new relevance assessments. The robust task reused existing CLEF test collections and did not create any new topics or make any fresh relevance assessments.

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

Language	Collections
Bulgarian	Sega 2002, Standart 2002, Novinar 2002
Czech	Mlada fronta DNES 2002, Lidové Noviny 2002
English	LA Times 2002
Hungarian	Magyar Hirlap 2002

Table 2. Test collections for the Robust task

Language	Collections
English	LA Times 94, Glasgow Herald 95
French	ATS (SDA) 94/95, Le Monde 94
Portuguese	Publico 94/95, Folha de Sao Paulo 94/95

Documents. The document collections used for the CLEF 2007 Ad Hoc tasks are part of the CLEF multilingual corpus of newspaper and news agency documents described in the Introduction to these Proceedings.

In the main stream monolingual and bilingual tasks, Bulgarian, Czech, Hungarian and English national newspapers for 2002 were used. Much of this data represented new additions to the CLEF multilingual comparable text corpora: Czech is a totally new language in the ad hoc track although it was introduced into the speech retrieval track last year; the Bulgarian collection was expanded with the addition of another national newspaper, and in order to have comparable data for English, we acquired a new American-English collection: Los Angeles Times 2002. Table 1 summarizes the collections used for each language.

The robust task used test collections containing news documents for the period 1994-1995 in three languages (English, French, and Portuguese) used in CLEF 2000 through CLEF 2006. Table 2 summarizes the collections used for each language.

Topics. Topics in the CLEF ad hoc track are structured statements representing information needs; the systems use the topics to derive their queries. Each topic consists of three parts: a brief “title” statement; a one-sentence “description”; a more complex “narrative” specifying the relevance assessment criteria.

Sets of 50 topics were created for the CLEF 2007 ad hoc mono- and bilingual tasks. All topic sets were created by native speakers. One of the decisions taken early on in the organization of the CLEF ad hoc tracks was that for each task the same set of topics, rendered in different languages, would be used to query the different collections. There were a number of reasons for this: it makes it easier to compare results over different target languages, 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. In CLEF 2006 we deviated from this rule as we were using document collections from two distinct periods (1994/5 and 2002) and created partially separate (but overlapping) sets with a common set of time-independent topics and separate sets of time-specific topics. As we had expected this really complicated our lives as we had to build more

topics and had to specify very carefully which topic sets were to be used against which document collections¹. We decided not to repeat this experience this year and thus only used collections from the same time period.

We created topics in both European and non-European languages. European language topics were offered for Bulgarian, Czech, English, French, Hungarian, Italian and Spanish. The non-European topics were prepared according to demand from participants. This year we had Amharic, Chinese, Indonesian, Oromo plus the group of Indian languages: Bengali, Hindi, Marathi, Tamil and Telugu.

The provision of topics in unfamiliar scripts did lead to some problems. These were not caused by encoding issues (all CLEF data is encoded using UTF-8) but rather by errors in the topic sets which were very difficult for us to spot. Although most such problems were quickly noted and corrected, and the participants were informed so that they all used the right set, one did escape our notice: the title of Topic 430 in the Czech set was corrupted and systems using Czech thus did not do well with this topic. It should be remembered, however, that an error in one topic does not really impact significantly on the comparative results of the systems. The topic will, however, be corrected for future use.

This year topics have been identified by means of a Digital Object Identifier (DOI)² of the experiment [1] which allows us to reference and cite them. Below we give an example of the English version of a typical CLEF 2007 topic:

```
<top lang="en">
<num>10.2452/401-AH</num>
<title>Euro Inflation</title>
<desc>Find documents about rises in prices after the introduction of the
Euro.</desc>
<narr>Any document is relevant that provides information on the rise of
prices in any country that introduced the common European
currency.</narr>
</top>
```

For the robust task, topic sets from CLEF 2001 to 2006 in English, French and Portuguese were used. For English and French, in CLEF for more time, training topics were offered and a set of 100 topics were used for testing. For Portuguese, no training topics were possible and a set of 150 test topics was used.

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,

¹ This is something that anyone reusing the CLEF 2006 ad hoc test collection needs to be very careful about.

² In order to resolve the DOIs used in this paper and to access on-line the relative information, you can use any DOI resolver, such as <http://dx.doi.org/>.

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 is 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 website³.

As a consequence of limited evaluation resources, a maximum of 12 runs each for the mono- and bilingual tasks was allowed (no more than 4 runs for any one language combination - we try to encourage diversity). For bi- and monolingual robust tasks, 4 runs were allowed per language or language pair.

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 selected runs according to a set of predefined criteria. Traditionally, the top 100 ranked documents from each of the runs selected are included in the pool; in such a case we say that the pool is of depth 100. This pool is then used for subsequent relevance judgments. After calculating the effectiveness measures, the results are analyzed and run statistics produced and distributed.

The stability of pools constructed in this way and their reliability for post-campaign experiments is discussed in [2] with respect to the CLEF 2003 pools. New pools were formed in CLEF 2007 for the runs submitted for the main stream mono- and bilingual tasks. Instead, the robust tasks used the original pools and relevance assessments from previous CLEF campaigns.

The main criteria used when constructing these pools were:

- favour diversity among approaches adopted by participants, according to the descriptions of the experiments provided by the participants;
- choose at least one experiment for each participant in each task, selected from the experiments with highest priority as indicated by the participant;
- add mandatory title+description experiments, even though they do not have high priority;
- add manual experiments, when provided;
- for bilingual tasks, ensure that each source topic language is represented.

³ http://trec.nist.gov/trec_eval/

One important limitation when forming the pools is the number of documents to be assessed. We estimate that assessors can judge from 60 to 100 documents per hour, providing binary judgments: relevant / not relevant. This is actually an optimistic estimate and shows what a time-consuming and resource expensive task human relevance assessment is. This limitation impacts strongly on the application of the criteria above - and implies that we are obliged to be flexible in the number of documents judged per selected run for individual pools.

This meant that this year, in order to create pools of more-or-less equivalent size (approx. 20,000 documents), the depth of the Bulgarian, Czech and Hungarian pools varied: 60 for Czech and 80 for Bulgarian and Hungarian, rather than the depth of 100 originally used to judge TREC ad hoc experiments⁴. In his paper in these proceedings, Tomlinson [3] makes some interesting observations in this respect. He claims that on average, the percentage of relevant items assessed was less than 60% for Czech, 70% for Bulgarian and 85% for Hungarian. However, as Tomlinson also points out, it has already been shown that test collections created in this way do normally provide reliable results, even if not all relevant documents are included in the pool.

When building the pool for English, in order to respect the above criteria and also to obtain a pool depth of 60, we had to include more than 25,000 documents. Even so, as can be seen from Table 3, it was impossible to include very many runs - just one monolingual and one bilingual run for each set of experiments.

The box plot of Figure 1 compares the distributions of the relevant documents across the topics of each pool for the different ad hoc pools; the boxes are ordered by decreasing mean number of relevant documents per topic. As can be noted, Bulgarian, Czech, and Hungarian distributions appear similar, even though the Czech and Hungarian ones are slightly more asymmetric towards topics with a greater number of relevant documents. On the other hand, the English distribution presents a greater number of relevant documents per topic, with respect to the other distributions, and is quite asymmetric towards topics with a greater number of relevant documents. All the distributions show some upper outliers, i.e. topics with a great number of relevant document with respect to the behaviour of the other topics in the distribution. These outliers are probably due to the fact that CLEF topics have to be able to retrieve relevant documents in all the collections; therefore, they may find considerably more relevant documents in one collection than in others depending on the contents of the separate datasets. Thus, typically, each pool will have a different set of outliers.

Table 3 reports summary information on the 2007 ad hoc pools used to calculate the results for the main monolingual and bilingual experiments. In particular, for each pool, we show the number of topics, the number of runs submitted, the number of runs included in the pool, the number of documents in the pool (relevant and non-relevant), and the number of assessors.

⁴ Tests made on NTCIR pools in previous years have suggested that a depth of 60 is normally adequate to create stable pools, presuming that a sufficient number of runs from different systems have been included.

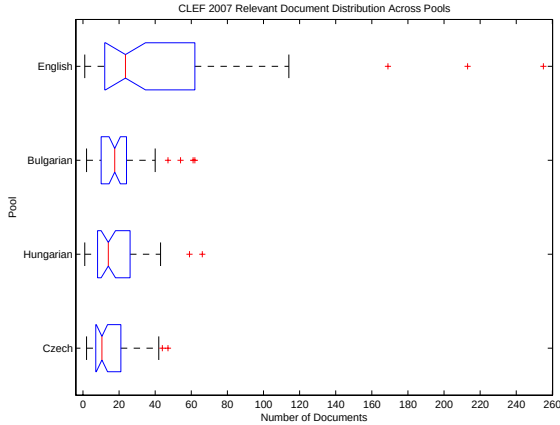


Fig. 1. Distribution of the relevant documents across the pools

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 [5]. For the robust task, we used different measures, see below Section 5.

The individual results for all official ad hoc experiments in CLEF 2007 are given in the Appendixes at the end of the Working Notes distributed for the workshop [6,7].

2.5 Participants and Experiments

22 groups from 12 different countries submitted results for one or more of the ad hoc tasks - a slight decrease on the 25 participants of last year. These groups submitted a total of 235 runs, a decrease of about 20% on the 296 runs of 2006. The average number of submitted runs per participant also slightly decreased: from 11.7 runs/participant of 2006 to 10.6 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 (138 out of 235, 58.72%) used this combination of topic fields, 50 (21.28%) used all fields, 46 (19.57%) used the title field, and only 1 (0.43%) used just the description field. The majority of experiments were conducted using automatic query construction (230 out of 235, 97.87%) and only in a small fraction of the experiments (5 out 237, 2.13%) were queries been manually constructed from topics. A breakdown into the separate tasks is shown in Table 4(a).

Table 3. Summary information about CLEF 2007 pools

Bulgarian Pool	
Pool size	19,441 pooled documents – 18,429 not relevant documents – 1,012 relevant documents 50 topics
Pooled Experiments	13 out of 18 submitted experiments – monolingual: 11 out of 16 submitted experiments – bilingual: 2 out of 2 submitted experiments
Assessors	4 assessors
Czech Pool	
Pool size	20,607 pooled documents – 19,485 not relevant documents – 762 relevant documents 50 topics
Pooled Experiments	19 out of 29 submitted experiments – monolingual: 17 out of 27 submitted experiments – bilingual: 2 out of 2 submitted experiments
Assessors	4 assessors
English Pool	
Pool size	24,855 pooled documents – 22,608 not relevant documents – 2,247 relevant documents 50 topics
Pooled Experiments	20 out of 104 submitted experiments – monolingual: 10 out of 31 submitted experiments – bilingual: 10 out of 73 submitted experiments
Assessors	5 assessors
Hungarian Pool	
Pool size	18,704 pooled documents – 17,793 not relevant documents – 911 relevant documents 50 topics
Pooled Experiments	14 out of 21 submitted experiments – monolingual: 12 out of 19 submitted experiments – bilingual: 2 out of 2 submitted experiments
Assessors	6 assessors

Table 4. Breakdown of experiments into tracks and topic languages

(a) Number of experiments per track, participant.

Track	# Part.	# Runs
Monolingual-BG	5	16
Monolingual-CS	8	27
Monolingual-EN	10	31
Monolingual-HU	6	19
Bilingual-X2BG	1	2
Bilingual-X2CS	1	2
Bilingual-X2EN	10	73
Bilingual-X2HU	1	2
Robust-Mono-EN	3	11
Robust-Mono-FR	5	12
Robust-Mono-PT	4	11
Robust-Bili-X2FR	3	9
Robust-Training-Mono-EN	2	6
Robust-Training-Mono-FR	2	6
Robust-Training-Bili-X2FR	2	8
Total		235

(b) List of experiments by topic language.

Topic Lang.	# Runs
English	73
Hungarian	33
Czech	26
Bulgarian	16
Indonesian	16
French	14
Hindi	13
Chinese	12
Portuguese	11
Amharic	9
Bengali	4
Oromo	4
Marathi	2
Telugu	2
Total	235

Fourteen different topic languages were used in the ad hoc experiments. As always, the most popular language for queries was English, with Hungarian second. The number of runs per topic language is shown in Table 4(b).

3 Main Stream Monolingual Experiments

Monolingual retrieval focused on central-European languages this year, with tasks offered for Bulgarian, Czech and Hungarian. Eight groups presented results for 1 or more of these languages. We also requested participants in the bilingual-to-English task to submit one English monolingual run, but only in order to provide a baseline for their bilingual experiments and in order to strengthen the English pool for relevance assessment⁵.

Five of the participating groups submitted runs for all three languages. One group was unable to complete its Bulgarian experiments, submitting results for just the other two languages. However, they subsequently completed their work on Bulgarian post-campaign, and results for all three languages are reported in this volume [11]. The two groups from the Czech Republic only submitted runs for Czech. From Table 5, it can be seen that the best performing groups were more-or-less the same for each language and that the results did not greatly differ. It should be noted that these are all veteran participants with much experience at CLEF.

⁵ Ten groups submitted runs for monolingual English. We have included a graph showing the top 5 results but it must be remembered that the systems submitting these were actually focusing on the bilingual part of the task.

As usual in the CLEF monolingual task, the main emphasis in the experiments was on stemming and morphological analysis. The group from University of Neuchatel, which had the best overall performances for all languages, focused very much on stemming strategies, testing both light and aggressive stemmers for the Slavic languages (Bulgarian and Czech). For Hungarian they worked on decompounding. This group also compared performances obtained using word-based and 4-gram indexing strategies [9]. Another of the best performers, from Johns Hopkins University, normally uses an n-gram approach. Unfortunately, we have not received a paper from this group so cannot comment on their performance. The other group with very good performance for all languages was Opentext. In their working notes paper, this group also compared 4-gram results against results using stemming for all three languages. They found that while there could be large impacts on individual topics, there was little overall difference in average performance. In addition, their experiments confirmed past findings that indicate that blind relevance feedback can be detrimental to results, depending on the evaluation measures used [4]. The results of the statistical tests that can be found towards the end of our working notes paper show that the best results of these three groups did not differ significantly [8].

The group from Alicante also achieved good results testing query expansion techniques [10], while the group from Kolkata compared a statistical stemmer against other types of stemmers for Bulgarian, Czech and Hungarian with comparable results [11]. Czech is a morphologically complex language and the two Czech-only groups both used approaches involving morphological analysis and lemmatization [12], [13].

3.1 Results

Table 5 shows 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 experiment; the DOI of the experiment; and the performance difference between the first and the last participant.

Figures comparing the performances of the top participants can be found in our working notes paper [8].

4 Main Stream Bilingual Experiments

The bilingual task was structured in three sub-tasks ($X \rightarrow BG, CS, \text{ or } HU$ target collection) plus a sub-task for non-European topic languages against an English target collection. A special sub-task testing Indian languages against the English collection was also organised in response to requests from a number of research groups working in India. For the bilingual to English task, participating groups also had to submit an English monolingual run, to be used both as baseline and also to reinforce the English pool. All groups participating in the Indian languages sub-task also had to submit at least one run in Hindi (mandatory) plus runs in other Indian languages (optional).

Table 5. Best entries for the monolingual track

Track	Rank	Participant	Experiment DOI	MAP
Bulgarian	1st	unine	10.2415/AH-MONO-BG-CLEF2007.UNINE.UNINEBG4	44.22%
	2nd	jhu-apl	10.2415/AH-MONO-BG-CLEF2007.JHU-APL.APLMOBGTD4	36.57%
	3rd	opentext	10.2415/AH-MONO-BG-CLEF2007.OPENTEXT.OTBG07TDE	35.02%
	4th	alicante	10.2415/AH-MONO-BG-CLEF2007.ALICANTE.IRNBUEXP2N	29.81%
	5th	daedalus	10.2415/AH-MONO-BG-CLEF2007.DAEDALUS.BGFSBG2S	27.19%
	Difference			62.33%
Czech	1st	unine	10.2415/AH-MONO-CS-CLEF2007.UNINE.UNINECZ4	42.42%
	2nd	jhu-apl	10.2415/AH-MONO-CS-CLEF2007.JHU-APL.APLMOCSSTD4	35.86%
	3rd	opentext	10.2415/AH-MONO-CS-CLEF2007.OPENTEXT.OTCS07TDE	34.84%
	4th	prague	10.2415/AH-MONO-CS-CLEF2007.PRAGUE.PRAGUE01	34.19%
	5th	daedalus	10.2415/AH-MONO-CS-CLEF2007.DAEDALUS.CSFSCS2S	32.03%
	Difference			32.44%
Hungarian	1st	unine	10.2415/AH-MONO-HU-CLEF2007.UNINE.UNINEHU4	47.73%
	2nd	opentext	10.2415/AH-MONO-HU-CLEF2007.OPENTEXT.OTHU07TDE	43.34%
	3rd	alicante	10.2415/AH-MONO-HU-CLEF2007.ALICANTE.IRNBUEXP2N	40.09%
	4th	jhu-apl	10.2415/AH-MONO-HU-CLEF2007.JHU-APL.APLMOHUTD5	39.91%
	5th	daedalus	10.2415/AH-MONO-HU-CLEF2007.DAEDALUS.HUFSHU2S	34.99%
	Difference			36.41%
English (only for Bilingual X2EN participants)	1st	bombay-ltrc	10.2415/AH-MONO-EN-CLEF2007.BOMBAY-LTRC.IITB_MONO_TITLE_DESC	44.02%
	2nd	jhu-apl	10.2415/AH-MONO-EN-CLEF2007.JHU-APL.APLMOENTD5	43.42%
	3rd	nottingham	10.2415/AH-MONO-EN-CLEF2007.NOTTINGHAM.MONOT	42.74%
	4th	depok	10.2415/AH-MONO-EN-CLEF2007.DEPOK.UIQTDMONO	40.57%
	5th	hyderabad	10.2415/AH-MONO-EN-CLEF2007.HYDERABAD.ENTD_QMENG07	40.16%
	Difference			9.61%

We were disappointed to only receive runs from one participant for the $X \rightarrow$ BG, CS, or HU tasks. Furthermore, the results were quite poor; as this group normally achieves very good performance, we suspect that these runs were probably corrupted in some way. For this reason, we decided to disregard them as being of little significance. Therefore, in the rest of this section, we only comment on the $X \rightarrow$ EN results.

We received runs using the following topic languages: Amharic, Chinese, Indonesian and Oromo plus, for the Indian sub-task, Bengali, Hindi, Marathi and Telugu⁶.

For many of these languages few processing tools or resources are available. It is thus very interesting to see what measures the participants adopted to overcome this problem. Here below, we briefly glance at some of the approaches and techniques adopted. For more details, see the papers cited.

The top performance in the bilingual task was obtained by an Indonesian group; they compared different translation techniques: machine translation using Internet resources, transitive translation using bilingual dictionaries and French and German as pivot languages, and lexicons derived from parallel corpus created by translating all the CLEF English documents into Indonesian using a commercial MT system. They found that they obtained best results using the MT system together with query expansion [14].

⁶ Although topics had also been requested in Tamil, in the end they were not used.

The second placed group used Chinese for their queries and a dictionary based translation technique. The experiments of this group concentrated on developing new strategies to address two well-known CLIR problems: translation ambiguity, and coverage of the lexicon [15]. The work by [16] which used Amharic as the topic language also paid attention to the problems of sense disambiguation and out-of-vocabulary terms. They found that pseudo-relevance feedback improved their performance considerably.

The third performing group also used Indonesian as the topic language; unfortunately we have not received a paper from them so far so cannot comment on their approach. An interesting paper, although slightly out of the task as the topic language used was Hungarian was [17]. This group used a machine readable dictionary approach but also applied Wikipedia hyperlinks for query term disambiguation and exploited bilingual Wikipedia articles for dictionary extension. The group testing Oromo used linguistic and lexical resources developed at their institute; they adopted a bilingual dictionary approach and also tested the impact of a light stemmer for Oromo on their performance with positive results [18].

The groups using Indian topic languages tested different approaches. The group from Kolkata submitted runs for Bengali, Hindi and Telugu to English using a bilingual dictionary lookup approach [19]. They had the best performance using Telugu probably because they carried out some manual tasks during indexing. A group from Bangalore tested a statistical MT system trained on parallel aligned sentences and a language modelling based retrieval algorithm for a Hindi to English system [20]. The group from Bombay had the best overall performances; they used bilingual dictionaries for both Hindi and Marathi to English and applied term-to-term cooccurrence statistics for sense disambiguation [21]. The Hyderabad group attempted to build bilingual dictionaries using topical similarity by choosing vocabulary from a web search engine index and demonstrated that such dictionaries perform very well even with fewer entries [18]. Interesting work was done by the group from Kharagpur which submitted runs for Hindi and Bengali. They attempted to overcome the lack of resources for Bengali by using phoneme-based transliterations to generate equivalent English queries from Hindi and Bengali topics [23].

4.1 Results

Table 6 shows the best results for the bilingual task. The performance difference between the best and the fifth 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 $\rightarrow$ EN.

For bilingual retrieval evaluation, a common method to evaluate performance is to compare results against monolingual baselines. This year we can only comment on the results for the bilingual to English tasks. The best results were obtained by a system using Indonesian as a topic language. This group achieved 88.10% of the best monolingual English IR system. This is a good result considering that Indonesian is not a language for which a lot of resources and machine-readable dictionaries are available. It is very close to the best results obtained

Table 6. Best entries for the bilingual task

Track	Rank	Part.	Lang.	Experiment DOI	MAP
Bulgarian	1st	jhu-apl	en	10.2415/AH-BILI-X2BG-CLEF2007.JHU-APL.APLBIENBGTd4	7.33%
	Difference				
Czech	1st	jhu-apl	en	10.2415/AH-BILI-X2CS-CLEF2007.JHU-APL.APLBIENCSTd4	21.43%
	Difference				
English	1st	depok	id	10.2415/AH-BILI-X2EN-CLEF2007.DEPOK.UIQTDTOGGLEFB10D10T	38.78%
	2nd	nottingham	zh	10.2415/AH-BILI-X2EN-CLEF2007.NOTTINGHAM.GRAWOTD	34.56%
	3rd	jhu-apl	id	10.2415/AH-BILI-X2EN-CLEF2007.JHU-APL.APLBIIDENTS	33.24%
	4th	hyderabad	om	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.OMTD07	29.91%
	5th	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	Difference				31.37%
Hungarian	1st	jhu-apl	en	10.2415/AH-BILI-X2HU-CLEF2007.JHU-APL.APLBIENHUTd5	29.63%
	Difference				

Table 7. Best entries for the bilingual Indian subtask

Track	Rank	Part.	Lang.	Experiment DOI	MAP
Hindi to English	1st	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	2nd	msindia	hi	10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_RBLM_ALL_CROSS_1000_POSSCORES	21.80%
	3rd	hyderabad	hi	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.HITD	15.60%
	4th	jadavpur	hi	10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILIH2ENR1	10.86%
	5th	kharagpur	hi	10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.HINDITITLE	4.77%
	6th				
	Difference				518.87%
Bengali/ Hindi/ Marathi/ Telugu to English	1st	bombay-ltrc	hi	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_HINDI_TITLEDESC_DICE	29.52%
	2nd	msindia	hi	10.2415/AH-BILI-X2EN-CLEF2007.MSINDIA.2007_RBLM_ALL_CROSS_1000_POSSCORES	21.80%
	3rd	bombay-ltrc	mr	10.2415/AH-BILI-X2EN-CLEF2007.BOMBAY-LTRC.IITB_MAR_TITLEDICE	21.63%
	4th	hyderabad	te	10.2415/AH-BILI-X2EN-CLEF2007.HYDERABAD.TETD	21.55%
	5th	jadavpur	te	10.2415/AH-BILI-X2EN-CLEF2007.JADAVPUR.AHBILITE2ENR1	11.28%
	6th	kharagpur	bn	10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.BENGALITITLEDISC	7.25%
	Difference				307.17%

last year for two well-established CLEF languages: French and Portuguese, when the equivalent figures were 93.82% and 90.91%, respectively.

4.2 Indian to English Subtask Results

Table 7 shows the best results for the Indian sub-task. The performance difference between the best and the last (up to 6) placed group is given (in terms of average precision). The first set of rows regards experiments for the mandatory topic language: Hindi; the second set of rows reports experiments where the source language is one of the other Indian languages.

It is interesting to note that in both sets of experiments, the best performing participant is the same. In the second set, we can note that for three (Hindi, Marathi, and Telugu) out of the four Indian languages used the performances of the top groups are quite similar.

The best performance for the Indian sub-task is 76.12% of the best bilingual English system (achieved by veteran CLEF participants) and 67.06% of the monolingual baseline, which is quite encouraging for a new task with languages where encoding issues and linguistic resources make the task difficult. This is in fact comparable with the performances of some newly introduced European languages. For example, we can compare them to those for Bulgarian and Hungarian in CLEF 2006:

- $X \rightarrow \text{BG}$: 52.49% of best monolingual Bulgarian IR system;
- $X \rightarrow \text{HU}$: 53.13% of best monolingual Hungarian IR system.

5 Robust Experiments

The robust task ran for the second time at CLEF 2007. It is an ad-hoc retrieval task based on test suites of previous CLEF campaigns. The evaluation approach is modified and a different perspective is taken. The robust task emphasizes the difficult topics by a non-linear integration of the results of individual topics into one result for a system [24,25]. By doing this, the evaluation results are interpreted in a more user-oriented manner. Failures and very low results for some topics hurt the user experience with a retrieval system. Consequently, any system should try to avoid these failures. This has turned out to be a hard task [26]. Robustness is a key issue for the transfer of research into applications. The robust task rewards systems which achieve a minimal performance level for all topics.

In order to do this, the robust task uses the geometric mean of the average precision for all topics (GMAP) instead of the mean average of all topics (MAP). This measure has also been used at a robust track at the Text Retrieval Conference (TREC) where robustness was explored for monolingual English retrieval [25]. At CLEF 2007, robustness was evaluated for monolingual and bilingual retrieval for three European languages.

The robust task at CLEF exploits data created for previous CLEF editions. Therefore, a test set with 100 topics can be used for the evaluation. Such a large number of topics allows a more reliable evaluation [27]. A secondary goal of the robust task is the definition of larger data sets for retrieval evaluation.

As described above, the CLEF 2007 robust task offered three languages often used in previous CLEF campaigns: English, French and Portuguese. The data used was developed during CLEF 2001 through 2006. Generally, the topics from CLEF 2001 until CLEF 2003 were used as training topics whereas the topics developed between 2004 and 2006 were the test topics on which the main evaluation measures are given.

Thus, the data used in the robust task in 2007 is different from the set defined for the robust task at CLEF 2006. The documents which need to be searched are articles from major newspapers and news providers in the three languages. Not all collections had been offered consistently for all CLEF campaigns, therefore, not all collections were integrated into the robust task. Most data from 1995 was omitted in order to provide a homogeneous collection. However, for Portuguese,

Table 8. Data for the Robust Task 2007

Language	Target Collection	Training Topic DOIs	Test Topic DOIs
English	LA Times 1994	10.2452/41-AH-10.2452/200-AH	10.2452/251-AH-10.2452/350-AH
French	Le Monde 1994 SDA 1994	10.2452/41-AH-10.2452/200-AH	10.2452/251-AH-10.2452/350-AH
Portuguese	Público 1995	—	10.2452/201-AH-10.2452/350-AH

Table 9. Best entries for the robust monolingual task

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	1st	reina	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.REINA.REINAENTDNT	38.97%	18.50%
	2nd	daedalus	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.DAEDALUS.EMFSNZ22S	37.78%	17.72%
	3rd	hildesheim	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2007.HILDESHEIM.HIMOENBRFNE	5.88%	0.32%
	4th				
	5th				
	Difference			562.76%	5,681.25%
French	1st	unine	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.UNINE.UNINEFR1	42.13%	14.24%
	2nd	reina	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.REINA.REINAFRTDET	38.04%	12.17%
	3rd	jaen	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.JAEN.UJARTFR1	34.76%	10.69%
	4th	daedalus	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.DAEDALUS.FRFSFR22S	29.91%	7.43%
	5th	hildesheim	10.2415/AH-ROBUST-MONO-FR-TEST-CLEF2007.HILDESHEIM.HIMOFRBRF2	27.31%	5.47%
	Difference			54.27%	160.33%
Portuguese	1st	reina	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.REINA.REINAPTTDNT	41.40%	12.87%
	2nd	jaen	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.JAEN.UJARTPT1	24.74%	0.58%
	3rd	daedalus	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.DAEDALUS.PTFSPT2S	23.75%	0.50%
	4th	xldb	10.2415/AH-ROBUST-MONO-PT-TEST-CLEF2007.XLDB.XLDBRBR16_10	1.21%	0.07%
	5th				
	Difference			3,321.49%	18,285.71%

for which no training data was available, only data from 1995 was used. Table 8 shows the data for the robust task.

The robust task attracted 63 runs submitted by seven groups (CLEF 2006: 133 runs from eight groups). Effectiveness scores were calculated with version 8.0 of the `trec_eval` program which provides the *Mean Average Precision (MAP)*, while the DIRECT system version 2.0 was used to calculate the *Geometric Average Precision (GMAP)*. .

5.1 Robust Monolingual Results

Table 9 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). The results cannot be compared to the results of the CLEF 2005 and CLEF 2006 campaign in which the same topics were used because a smaller collection had to be searched. The working notes paper contains figures comparing the performances of the top participants for each language [8].

Table 10. Best entries for the robust bilingual task

Track	Rank	Participant	Experiment DOI	MAP	GMAP
French	1st	reina	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.REINA.REINAE2FTDNT	35.83%	12.28%
	2nd	unine	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.UNINE.UNINEBILFR1	33.50%	5.01%
	3rd	colesun	10.2415/AH-ROBUST-BILI-X2FR-TEST-CLEF2007.COLESUN.EN2FRTST4GRINTLOGLU001	22.87%	3.57%
	4th				
	5th				
	Difference			54.27%	243.98%

5.2 Robust Bilingual Results

Table 10 shows the best results for this task. The performance difference between the best and the last (up to 5) placed group is given (in terms of average precision). All the experiments were from English to French.

As previously stated for bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We have the following results for CLEF 2007:

- $X \rightarrow FR$: 85.05% of best monolingual French IR system;

The results in Table 9 and Table 10 suggest that there is no difference between the rankings based on MAP and those based on GMAP. No position changes between system occur. However, a more thorough analysis of the CLEF 2006 robust results which included not only the best run of each system but also the other runs showed interesting results. The correlation between MAP and GMAP based rankings is decreasing with the number of topics for multilingual retrieval [32]. This result shows that the creation of larger test suites is necessary.

5.3 Approaches Applied to Robust Retrieval

The REINA system performed best in three tasks and seems to be well optimized for robust retrieval [34]. It applied different measures of robustness during the training phase in order to optimize the performance. A local query expansion technique added terms. Another system experimented with n-gram based translation for bi-lingual retrieval which requires no languages specific components [36]. SINAI tried to increase the robustness of the results by expanding the query with an external knowledge source [33]. This is a typical approach in order to obtain additional query terms and avoid zero hits in case of out of vocabulary problems. Contrary to standard query expansion techniques, the new terms form a second query and results of both initial and second query are integrated under a logistic fusion strategy. The Daedalus group submitted experiments with the Miracle system [35]; BM25 weighting without blind relevance feedback was applied. For detailed descriptions of all the robust experiments, see the Robust section in these Proceedings.

6 Conclusions

We have reported the results of the ad hoc cross-language textual document retrieval track at CLEF 2007. 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 text retrieval systems and compare performance between monolingual and cross-language runs, before perhaps moving on to more complex system development and subsequent evaluation. This year, the monolingual task focused on central European languages while the bilingual task included an activity for groups that wanted to use non-European topic languages and languages with few processing tools and resources. Each year, we also include a task aimed at examining particular aspects of cross-language text retrieval. Again this year, the focus was examining the impact of “hard” topics on performance in the “robust” task.

The paper also describes in some detail the creation of the pools used for relevance assessment this year and includes observations on their stability. We also performed a number of statistical tests on the results with the aim of determining what differences between runs appear to be real as opposed to differences that are due to sampling issues. Unfortunately, for reasons of space, we are unable to report the results here. The interested reader is again referred to our on-line working notes paper.

Although there was quite a good participation in the monolingual Bulgarian, Czech and Hungarian tasks and the experiments report some interesting work on stemming and morphological analysis, we were very disappointed by the lack of participation in bilingual tasks for these languages. On the other hand, the interest in the task for non-European topic languages was encouraging and the results reported can be considered positively.

The robust task has analyzed the performance of systems for older CLEF data under a new perspective. A larger data set which allows a more reliable comparative analysis of systems was assembled. Systems needed to avoid low performing topics. Their success was measured with the geometric mean (GMAP) which introduces a bias on poor performing topics. Results for the robust task for mono-lingual retrieval on English, French and Portuguese collections as well as for bi-lingual retrieval from English to French are reported. Robustness can also be interpreted as the fitness of a system under a variety of conditions. The definition on what robust retrieval means has to continue.

As a result of discussions at the workshop, the CLEF 2008 Ad Hoc track has been considerably revolutionized. We offer a totally new main task for monolingual and cross-language search on library catalogue records, organised in collaboration with The European Library (TEL). We also offer more traditional mono- and bilingual ad-hoc retrieval tasks on a Persian newspaper corpus: the Hamshahri collection. This is the first time that we offer a non-European target collection in CLEF. The 2008 “robust” task proposes monolingual and bilingual tasks on a word sense disambiguated (WSD) collection of news documents in English. The goal of the task is to test whether WSD can be used beneficially for retrieval systems. The results will be reported and discussed at the CLEF 2008 workshop.

Acknowledgements

We should like to acknowledge the enormous contribution of the groups responsible for topic creation and relevance assessment. In particular, we thank the group responsible for the work on Bulgarian led by Kiril Simov and Petya Osenova, the group responsible for Czech led by Pavel Pecina⁷, and the group responsible for Hungarian led by Tamás Váradi and Gergely Bottyán. These groups worked very hard under great pressure in order to complete the heavy load of relevance assessments in time.

References

1. Paskin, N. (ed.): The DOI Handbook – Edition 4.4.1. International DOI Foundation (IDF) (2006) [last visited 2007, August 30], <http://dx.doi.org/10.1000/186>
2. Braschler, M.: CLEF 2003 - Overview of results. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 44–63. Springer, Heidelberg (2004)
3. Tomlinson, S.: Sampling Precision to Depth 10000 at CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 57–64. Springer, Heidelberg (2008)
4. Tomlinson, S.: Sampling Precision to Depth 10000: Evaluation Experiments at CLEF 2007. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007) [last visited May 2008], <http://www.clef-campaign.org/>
5. 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)
6. Di Nunzio, G.M., Ferro, N.: Appendix A: Results of the Core Tracks – Ad-hoc Bilingual and Monolingual Tasks. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007) [last visited May 2008], <http://www.clef-campaign.org/>
7. Di Nunzio, G.M., Ferro, N.: Appendix B: Results of the Core Tracks – Ad-hoc Robust Bilingual and Monolingual Tasks. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007) [last visited May 2008], <http://www.clef-campaign.org/>
8. Di Nunzio, G.M., Ferro, N., Peters, C., Mandl, T.: CLEF 2007: Ad Hoc Track Overview. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop [last visited May 2008], <http://www.clef-campaign.org/>
9. Dolamic, L., Savoy, J.: Stemming Approaches for East European Languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 37–44. Springer, Heidelberg (2008)
10. Noguera, E., Llopis, F.: Applying Query Expansion Techniques to Ad Hoc Monolingual Tasks with the IR-n system. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 45–48. Springer, Heidelberg (2008)
11. Majumder, P., Mitra, M., Pal, D.: Bulgarian, Hungarian and Czech Stemming using YASS. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 49–56. Springer, Heidelberg (2008)
12. Češka, P., Pecina, P.: Charles University at CLEF 2007 Ad-Hoc Track. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 33–36. Springer, Heidelberg (2008)

⁷ Ministry of Education of the Czech Republic, project MSM 0021620838.

13. Ircing, P., Müller, L.: Czech Monolingual Information Retrieval Using Off-The-Shelf Components – the University of West Bohemia at CLEF 2007 Ad-Hoc track. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007) [last visited May 2008], <http://www.clef-campaign.org/>
14. Adriani, M., Hayurani, H., Sari, S.: Indonesian-English Transitive Translation for Cross-Language Information Retrieval. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 127–133. Springer, Heidelberg (2008)
15. Zhou, D., Truran, M., Brailsford, T.: Disambiguation and Unknown Term Translation in Cross Language Information Retrieval. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 64–71. Springer, Heidelberg (2008)
16. Argaw, A.A.: Amharic-English Information Retrieval with Pseudo Relevance Feedback. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 119–126. Springer, Heidelberg (2008)
17. Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K.: Cross-Language Retrieval with Wikipedia. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 72–79. Springer, Heidelberg (2008)
18. Pingali, P., Tune, K.K., Varma, V.: Improving Recall for Hindi, telugu, Oromo to English CLIR. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 103–110. Springer, Heidelberg (2008)
19. Bandyopadhyay, S., Mondal, T., Naskar, S.K., Ekbal, A., Haque, R., Godavarthy, S.R.: Bengali, Hindi and Telugu to English Ad-hoc Bilingual task at CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 88–94. Springer, Heidelberg (2008)
20. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 80–87. Springer, Heidelberg (2008)
21. Chinnakotla, M.K., Ranadive, S., Damani, O.P., Bhattacharyya, P.: Hindi-English and Marathi-English Cross Language Information Retrieval Evaluation. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 111–118. Springer, Heidelberg (2008)
22. Pingali, P., Varma, V.: IIIT Hyderabad at CLEF 2007 – Adhoc Indian Language CLIR task. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop [last visited May 2008] <http://www.clef-campaign.org/>
23. Mandal, D., Gupta, M., Dandapat, S., Banerjee, P., Sarkar, S.: Bengali and Hindi to English CLIR Evaluation. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 95–103. Springer, Heidelberg (2008)
24. Robertson, S.: On GMAP: and Other Transformations. In: Yu, P.S., Tsotras, V., Fox, E.A., Liu, C.B. (eds.) Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), pp. 78–83. ACM Press, New York (2006)
25. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum 39, 11–20 (2005)
26. Savoy, J.: Why do Successful Search Systems Fail for Some Topics. In: Cho, Y., Wan Koo, Y., Wainwright, R.L., Haddad, H.M., Shin, S.Y. (eds.) Proc. 2007 ACM Symposium on Applied Computing (SAC 2007), pp. 872–877. ACM Press, New York (2007)
27. Sanderson, M., Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In: Baeza-Yates, R., Ziviani, N., Marchionini, G., Moffat, A., Tait, J. (eds.) Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005), pp. 162–169. ACM Press, New York (2005)

28. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: Korfthage, 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)
29. Conover, W.J.: *Practical Nonparametric Statistics*, 1st edn. John Wiley and Sons, New York (1971)
30. 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)
31. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. In: Spack Jones, K., Willett, P. (eds.) *Readings in Information Retrieval*, pp. 205–216. Morgan Kaufmann Publisher, Inc., San Francisco (1997)
32. Mandl, T., Womser-Hacker, C., Ferro, N., Di Nunzio, G.: How Robust are Multilingual Information Retrieval Systems? In: *Proc. 2008 ACM SAC Symposium on Applied Computing (SAC)*, pp. 1132–1136. ACM Press, New York (2008)
33. Martínez-Santiago, F., Ráez, A.M., Garcia Cumberras, M.A.: SINAI at CLEF Ad-Hoc Robust Track 2007: Applying Google Search Engine for Robust Cross-lingual Retrieval. In: Peters, C., et al. (eds.) *CLEF 2007. LNCS*, vol. 5152, pp. 137–142. Springer, Heidelberg (2008)
34. Zazo, A., Berrocal, J.L.A., Figuerola, C.G.: Improving Robustness Using Query Expansion. In: Peters, C., et al. (eds.) *CLEF 2007. LNCS*, vol. 5152, pp. 143–147. Springer, Heidelberg (2008)
35. González-Cristóbal, J.-C., Goñi-Menoyo, J.M., Goñi-Menoyo, J., Lana-Serrano, S.: MIRACLE Progress in Monolingual Information Retrieval at Ad-Hoc CLEF 2007. In: Peters, C., et al. (eds.) *CLEF 2007. LNCS*, vol. 5152, pp. 156–159. Springer, Heidelberg (2008)
36. Vilares, J., Oakes, M., Vilares Ferro, M.: English-to-French CLIR: A Knowledge-Light Approach through Character N- Grams Alignment. In: Peters, C., et al. (eds.) *CLEF 2007. LNCS*, vol. 5152, pp. 148–155. Springer, Heidelberg (2008)

Charles University at CLEF 2007 Ad-Hoc Track

Pavel Česka and Pavel Pecina

Charles University, Institute of Formal and Applied Linguistic
Malostranské náměstí 25, 118 00 Praha, Czech Republic
`{ceska,pecina}@ufal.mff.cuni.cz`

Abstract. In this paper we describe retrieval experiments performed at Charles University in Prague for participation in the CLEF 2007 Ad-Hoc track. We focused on the Czech monolingual task and used the LEMUR toolkit as the retrieval system. Our results demonstrate that for Czech as a highly inflectional language, lemmatization significantly improves retrieval results and manually created queries are only slightly better than queries automatically generated from topic specifications.

1 Introduction

This work represents the first participation of Charles University in the CLEF evaluation campaign. Our research is focused on Czech monolingual tasks and the application of advanced language processing tools developed at our university - namely a morphological analyser and tagger. We also attempt to compare systems with manually and automatically created queries. For the Ad-Hoc track we submitted four experiments (runs): *Prague01*, *Prague02*, *Prague03*, and *Prague04*. Our main goal were to study influence of lemmatization and whether manual query construction can bring additional performance improvement. Similar experiments were performed also for the CLEF 2007 Cross-Language Speech Retrieval track.

2 System Description

2.1 Retrieval Model

Being novices in the field of information retrieval we decided to use a freely available retrieval toolkit instead of developing our own. The final choice was the LEMUR toolkit [1] and its Indri retrieval model [2]. It is based on a combination of language modeling and inference network retrieval. A detailed description can be found in [3]. It has been popular among CLEF participant in recent years and was found effective for a wide range of retrieval tasks.

To improve retrieval results, we used Indri's pseudo-relevance feedback which is an adaption of Lawrenko's relevance models [4].

2.2 Morphological Tagging and Lemmatization

State-of-the-art retrieval systems usually include at least some basic linguistically-motivated preprocessing of the documents and queries such as stemming and stopword removal. Czech is a morphologically complex language and there is no easy way how to determine stems and their endings as it can be done in English and other languages. Stemming in Czech is not sufficient and should be replaced by a proper lemmatization (substituting each word by its base form – the lemma) which involves determining the part of speech of all words. In our experiments, we employed the Czech morphological analyzer and tagger developed at Charles University [5], [6] which assigns a disambiguated lemma and a morphological tag to each word. Its accuracy is around 95%. An example of its output for one word (“serious” in English) is following:

```
<f>závažných<MDl src="a">závažný<MDt src="a">AAlP2----1A----
```

The tag `<f>` is followed by the original word form, tag `<MDl>` is followed by the lemma, and the tag `<MDt>` separates a 15-position morphological category (the first position represents the part-of-speech; A stands for an adjective). Lemmatization was employed in all our experiments except *Prague03*. In *Prague01*, both original word forms and lemmas were used for indexing (in two separate model representations).

2.3 Stopword List Construction

We used two approaches to construct the stopword lists for our experiments. The first was based on the frequency of word occurrences in the collection, the latter on part-of-speech of words. In the first three experiments (*Prague01-03*), we removed the 40 most frequented words (separately from the original and lemmatized text) from the documents and the queries. In the fourth experiment (*Prague04*), we removed all words tagged as pronouns, prepositions, conjunctions, particles, interjections, and unknown words (mostly typos) and kept only open-class words.

2.4 Query Construction

Automatically created queries were constructed from the `<title>` and `<desc>` fields of the topic specifications only. The text was simply concatenated and processed by the analyzer and tagger. A combination of the original and lemmatized query was used in the first experiment (*Prague01*). Lemmatized queries containing only nouns, adjectives, numerals, adverbs and verbs were created for the fourth experiment (*Prague04*).

The queries in two of our experiments were created manually. In *Prague02* they were constructed from lemmas (to match the lemmatized documents) and their synonyms and in *Prague03* with the use of “stems“ and wildcard operators to cover all possible word forms (documents indexed in the original forms).

Example

The original title and description (topic 10.2452/413-AH: Reducing Diabetes Risk):

```
<title>Snižování rizika onemocnění cukrovkou</title>
<desc>Najděte dokumenty zmiňující faktory, které snižují riziko
onemocnění cukrovkou.</desc>
```

The *Prague01* query (original word forms plus lemmas; the suffixes *.(orig)* and *.(lemma)* refer to the corresponding model representations):

```
#combine(snižování.(orig) rizika.(orig) onemocnění.(orig)
cukrovkou.(orig) najděte.(orig) dokumenty.(orig) zmiňující.(orig)
faktory.(orig) které.(orig) snižují.(orig) riziko.(orig)
onemocnění.(orig) cukrovkou.(orig) snižování.(lemma) riziko.(lemma)
onemocnění.(lemma) cukrovka.(lemma) najít.(lemma) dokument.(lemma)
zmiňující.(lemma) faktor.(lemma) kter.(lemma) snižovat.(lemma)
riziko.(lemma) onemocnění.(lemma) cukrovka.(lemma))
```

The *Prague02* query based on lemmas (the operator *#combine()* combines beliefs of the nested operators, operator *#syn()* resets synonymic line of equal expressions and operator *#2()* represents ordered window with width 2 words):

```
#combine(#syn(diabetes cukrovka úplavice) #2(snížení riziko)
prevence)
```

The *Prague03* query with wildcard operators (which can be used as a suffix only).

```
#combine(diabet* cukrovk* úplavic* sníž* rizik* preven*)
```

The *Prague04* query:

```
#combine(snižování riziko onemocnění cukrovka zmiňující faktor snižovat
riziko onemocnění cukrovka)
```

3 Experiment Specification

The following table summarizes four experiment specifications which we submitted for the Ad-Hoc track.

	<i>Prague01</i>	<i>Prague02</i>	<i>Prague03</i>	<i>Prague04</i>
Topic fields	TD	TD	TD	TD
Query construction	<i>automatic</i>	<i>manual</i>	<i>manual</i>	<i>automatic</i>
Document fields	<title> <heading> <text>	<title> <heading> <text>	<title> <heading> <text>	<title> <heading> <text>
Word forms	<i>original+lemmas</i>	<i>lemmas</i>	<i>original</i>	<i>lemmas</i>
Stop words	<i>original+lemmas</i>	<i>lemmas</i>	<i>original</i>	<i>closed-class words</i>

4 Results and Conclusion

The Czech Ad-Hoc collection consists of 81,735 documents and 50 topics. The following table summarizes the results for the experiments described above.

	<i>Prague01</i>	<i>Prague02</i>	<i>Prague03</i>	<i>Prague04</i>
Mean Average Precision	0.3419	0.3336	0.3202	0.2969
Mean R Precision	0.3201	0.3349	0.3147	0.2886
Mean Binary Preference	0.2977	0.3022	0.2801	0.2601
Precision at 10 interpolated recall level	0.5733	0.6314	0.5299	0.5367

In terms of Mean Average Precision, the best score was achieved in experiment *Prague01*. Indexing both original word forms and lemmas in combination with automatically generated queries seems to be a reasonable way how to build a retrieval system. In terms of other performance measures, the scores of *Prague02* are slightly better but this is probably due to the use of synonyms in the manually created queries – not in the manual approach itself.

By comparing scores of *Prague03* with results of *Prague01* and *Prague02* we can confirm that lemmatization is quite useful for searching in highly flectional languages such a Czech and can not be fully substituted by stemming.

The last lesson we learned is that using extensive stopword lists based on part-of-speech can seriously harm the performance of a retrieval system as can be seen on the results of experiment *Prague04*.

We found these results quite encouraging and motivating for our future work.

Acknowledgments

This work has been supported by the Ministry of Education of the Czech Republic, projects MSM 0021620838 and #1P05ME786.

References

1. Lemur, <http://www.lemurproject.org/>
2. Indri, <http://www.lemurproject.org/indri/>
3. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language-model based search engine for complex queries (extended version). Technical Report IR-407, CIIR, UMass (2005)
4. Lavrenko, V., Croft, W.B.: Relevance based language models. In: SIGIR 2001: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York (2001)
5. Hajič, J., Vidová-Hladká, B.: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: Proceedings of the Conference COLING - ACL 1998, Montreal, Canada (1998)
6. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech), Nakladatelství Karolinum, Prague (2004)

Stemming Approaches for East European Languages

Ljiljana Dolamic and Jacques Savoy

Computer Science Department, University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchatel, Switzerland
Ljiljana.Dolamic@unine.ch, Jacques.Savoy@unine.ch

Abstract. During this CLEF evaluation campaign, the first objective is to propose and evaluate various indexing and search strategies for the Czech language that will hopefully result in more effective retrieval than language-independent approaches (n -gram). Based on the stemming strategy we developed for other languages, we propose that for the Slavic language a light stemmer (inflectional only) and also a second one based on a more aggressive suffix-stripping scheme that will remove some derivational suffixes. Our second objective is to undertake further study of the relative merit of various search engines when exploring Hungarian and Bulgarian documents. To evaluate these solutions we use various effective IR models. Our experiments generally show that for the Bulgarian language, removing certain frequently used derivational suffixes may improve mean average precision. For the Hungarian corpus, applying an automatic compounding procedure improves the MAP. For the Czech language a comparison of a light and a more aggressive stemmer to remove both inflectional and some derivational suffixes, reveals only small performance differences. For this language only, performance differences between a word-based or a 4-gram indexing strategy are also rather small.

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], [2] and popular Asian [3] languages (namely, Chinese, Japanese, and Korean). The main objective of our work has been to promote effective monolingual IR in these languages. For our participation in the CLEF 2007 evaluation campaign we thus decided to revamp our stemming strategy by including certain very frequently used derivational suffixes. When defining our stemming rules however we still focus on nouns and adjectives only. A description of the test-collections can be found in [4].

The rest of this paper is organized as follows: Section 2 outlines the main aspects of our stopword lists and stemming procedures. Section 3 analyses the principal features of different indexing and search strategies while Section 4 evaluates their use with the available corpora. Finally, Section 5 exposes our official results and Section 6 depicts our main findings.

2 Stemming Procedures

For the Hungarian language our suggested stemmer [5] mainly involves inflectional removal (gender, number and 23 grammatical cases, as for example in “házakat” → “ház” (house)) and also some pronouns (e.g., “házamat” (my house) → “ház”) and a few derivational suffixes (e.g., “temetés” (burial) → “temet” (to bury)). Because the Hungarian language uses compound constructions (e.g., “hétvége” (weekend) = “hét” (week / seven) + “vég” (end)), we increase matching possibilities between search keywords and document representations by automatically decompounded Hungarian words. To do so we apply our decompounding algorithm, leaving both compound words and their component parts in documents and queries. All stopword lists (containing 737 Hungarian forms) and stemmers used in this experiment are freely available at www.unine.ch/info/clef.

For the Bulgarian language we decided to modify the transliteration procedure we used previously to convert Cyrillic characters into Latin letters. We also modified last year’s stemmer, denoted as the light Bulgarian stemmer, by correcting an error and adapting it for the new transliteration scheme [2]. In this language, definite articles and plural forms are represented by suffixes and the general noun pattern is as follows: <stem> <plural> <article>. Our light stemmer contains eight rules for removing plurals and five for removing articles. Additionally we applied seven grammatical normalization rules plus three others to remove palatalization (changing stem’s final consonant when followed by a suffix beginning with certain vowels), as is very common in most Slavic languages. We also proposed a new and more aggressive Bulgarian stemmer that removes some derivational suffixes (e.g., “страшен” (fearful) → “страх” (fear)). The stopword list used for this language contains 309 words, somewhat bigger than that of last year (258 items).

For the Czech language, we proposed a new stopword list containing 467 forms (determinants, prepositions, conjunctions, pronouns, and some very frequent verb forms). We also designed and implemented two Czech stemmers. The first one is a light stemmer that removes only those inflectional suffixes attached to nouns or adjectives, in order to conflate to the same stem those morphological variations related to gender (feminine, neutral vs. masculine), number (plural vs. singular) and various grammatical cases (seven in the Czech language). For example, the noun “město” (city) appears as such in its singular form (nominative, vocative or accusative) but varies with other cases, “města” (genitive), “městu” (dative), “městem” (instrumental) or “městě” (locative). The corresponding plural forms are “města”, “měst”, “městům”, “městy” or “městech”. In the Czech language all nouns have a gender, and with a few exceptions (indeclinable borrowed words), they are declined for both number and case. For Czech nouns, the general pattern is as follows: <stem> <possessive> <case> in which <case> ending includes both gender and number. Adjectives are declined to match the gender, case and number of nouns to which they are attached. To remove these various case endings from nouns and adjectives we devised 52 rules,

and then before returning the computed stem, we added five normalization rules that control palatalization and certain vowel changes in the basic stem.

Finally, we designed and implemented a more aggressive stemmer that includes certain rules to remove frequently used derivational suffixes (e.g., “členstv” (membership) → “člen” (member)). In applying this second more aggressive stemmer (denoted “derivational”) we hope to improve mean average precision (MAP). Finally and unlike other languages, we do not remove the diacritic characters when building Czech stemmers.

3 Indexing and Searching Strategies

In order to obtain high MAP values, we considered adopting different weighting schemes for terms occurring in the documents or in the query. With this weighting we could account for term occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i), as well as their inverse document frequency (denoted idf_j). Moreover, we also considered normalize each indexing weight, using the cosine to obtain the classical $tf \cdot idf$ formulation.

In addition to this vector-space approach, we considered probabilistic models such as the Okapi [6] (or BM25). As a second probabilistic approach, we implemented three variants of the DFR (*Divergence from Randomness*) family of models suggested by Amati & van Rijsbergen [7]. Within this framework, indexing weights w_{ij} attached to term t_j in document D_i combine two information measures, expressed as follows:

$$w_{ij} = Inf_{ij}^1 \cdot Inf_{ij}^2 = -\log_2 [Prob_{ij}^1(tf)] \cdot (1 - Prob_{ij}^2) \quad (1)$$

As a first model, we implemented the GL2 scheme, defined as:

$$Prob_{ij}^1 = \left[\frac{1}{1 + \lambda_j} \right] \cdot \left[\frac{\lambda_j}{1 + \lambda_j} \right]^{tf_{n_{ij}}} \quad \text{with } \lambda_j = \frac{tc_j}{n} \quad (2)$$

$$Prob_{ij}^2 = \frac{tf_{n_{ij}}}{tf_{n_{ij}} + 1} \quad \text{with } tf_{n_{ij}} = tf_{ij} \cdot -\log_2 \left[1 + \frac{c \cdot \text{mean } dl}{l_i} \right] \quad (3)$$

where df_j indicates the number of documents in which term t_j occurs, tc_j the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i , *mean dl* the average document length, n the number of documents in the corpus, and c a constant.

As a second model, we implemented the PB2 scheme, defined as:

$$Inf_{ij}^1 = -\log_2 \left[\frac{e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}}}{tf_{ij}!} \right] \quad (4)$$

$$Prob_{ij}^2 = 1 - \left[\frac{tc_j + 1}{df_j \cdot (tf_{n_{ij}} + 1)} \right] \quad (5)$$

We then implemented a third model called IneC2 as follows:

$$Inf_{ij}^1 = tf_{n_{ij}} \cdot \left[\frac{n + 1}{n_e + 0.5} \right] \quad \text{with } n_e = n \cdot \left[1 - \left(\frac{n - 1}{n} \right)^{tc_j} \right] \quad (6)$$

$$Prob_{ij}^2 = 1 - \left[\frac{tc_j + 1}{df_j \cdot (tf n_{ij} + 1)} \right] \quad (7)$$

Finally, we considered an approach known as a non-parametric probabilistic model, based on a statistical language model (LM) [8]. As such, probability estimates would not be based on any known distribution (e.g., as in Equation 2), but rather be estimated directly, based on occurrence frequencies in document D_i or corpus C . Within this language model paradigm, various implementation and smoothing methods could be considered, although in this study we adopted a model proposed by Hiemstra [8], as described in Equation 8, combining an estimate based on document ($P[t_j|D_i]$) and on corpus ($P[t_j|C]$).

$$Prob[D_i|Q] = Prob[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot Prob[t_j|D_i] + (1 - \lambda_j) \cdot Prob[t_j|C]] \quad (8)$$

$$Prob[t_j|D_i] = tf_{ij}/l_i \quad \text{and} \quad Prob[t_j|C] = df_j/lc \quad \text{with} \quad lc = \sum_k df_k \quad (9)$$

where λ_j is a smoothing factor (constant for all indexing terms t_j , and fixed at 0.35) and lc an estimate of the size of the corpus C .

4 Evaluation

To measure the retrieval performance, we chose to use the mean average precision (MAP) obtained from 50 queries. In the following tables, the best performances under a given condition are listed in bold type. We then applied the bootstrap methodology [9] in order to statistically determine whether or not a given search strategy would be better than the performance depicted in bold. Thus, in the tables included in this paper we added an asterisk to indicate any statistically significant differences resulting from the use of a two-sided non-parametric bootstrap test ($\alpha = 5\%$).

Table 1 shows the MAP achieved by various probabilistic models using the Hungarian and Bulgarian collection, along with two different stemmers. An analysis of this data shows that the best performing IR model corresponds to the

Table 1. Evaluation of Hungarian and Bulgarian corpora

Query Stemmer	Mean average precision			
	Hungarian TD light	Hungarian TD + decomp.	Bulgarian TD light	Bulgarian TD derivat.
Okapi	0.3231*	0.3629*	0.3155*	0.3425*
DFR-GL2	0.3324*	0.3615*	0.3307	0.3541
DFR-IneC2	0.3525	0.3897	0.3423	0.3606
LM	0.3118*	0.3482*	0.3175*	0.3368*
<i>tf idf</i>	0.2344*	0.2532*	0.2103*	0.2143*

DFR-IneC2 model, with all stemming approaches and for both languages. For the Hungarian language, the best indexing strategy seems to be a word-based approach along with an automatic decompounding procedure. Using this strategy as a baseline, the average performance difference with an indexing strategy without a decompounding procedure is around 13% (DFR-IneC2: 0.3525 vs. 0.3897).

The evaluations done on the Czech language are depicted in Table 2. In this case, we compared two stemmers (light vs. derivational) and the 4-gram indexing approach (without stemming) [10]. The best performing IR model type is the DFR-IneC2 but the performance differences between the two DFR models are usually small. In the third column (labeled “no accent”) we evaluated the light stemmer, with all diacritic characters removed, and thus slightly reduced retrieval performance. When comparing the stemmers, the best indexing strategy seem to be the word-based indexing strategy, using the light stemming approach. Moreover, the performance differences between the 4-gram and this light stemming approach seem to be statistically not significant.

Table 2. Evaluation of the Czech Corpus

Query Stemmer	Mean average precision			
	TD light	TD no accent	TD derivat.	TD 4-grams
Okapi	0.3355	0.3306*	0.3255*	0.3401*
DFR-GL2	0.3437	0.3359	0.3342	0.3365
DFR-IneC2	0.3539	0.3473	0.3437	0.3517
LM	0.3263*	0.3174*	0.3109*	0.3304*
<i>tf idf</i>	0.2050*	0.2078*	0.1984*	0.2126*

A query-by-query analysis reveals that our various search strategies encountered some serious problems. For example with the Hungarian corpus, Topic #436 “VIP divorces” resulted in an average precision of 0.0003 because the term “VIP” is unknown in the collection and thus the query is composed of only a single and frequent word. With the Bulgarian corpus, Topic #429 “Water Health Risks” can be used to show the difference between our two stemming strategies. The search term “Health” is translated as “здравето” in the topic’s title, and we found the following forms in the relevant documents: “здравен”, “здравна” or “здравното”. When using our derivational stemmer, all these forms were conflated to the same stem (“здрав”) which was also the same stem for the word appearing in the query. With the light stemmer, the forms used in the relevant document were indexed under “здравн” which differs from the form appearing in the query (“здрав”). For the Czech corpus, we encountered a problem with spelling variations. With Topic #411 “Best picture Oscar”, the award name appears with two distinct spellings. In the Czech query however, the form used was “Oskar” (with a “k”) while in the relevant documents we found the form “Oscar”. The different search models were not able to find a match for the two forms.

Table 3. MAP Before and After Blind-Query Expansion

Query TD Stemmer	Mean average precision			
	Hungarian decompound	Hungarian decompound	Bulgarian derivation.	Czech light
Model Before	IneC2 0.3897	Okapi 0.3629	LM 0.3368	Okapi 0.3355
k docs/ m terms	5/20 0.4193*	5/20 0.3909*	10/50 0.4098*	5/20 0.3557*
	5/50 0.4284*	5/50 0.3973*	10/80 0.4043*	5/50 0.3610*
	5/70 0.4283*	5/70 0.3983*	10/100 0.4061*	5/70 0.3702*
	5/100 0.4298*	5/100 0.4010*	10/120 0.4004*	5/100 0.3685*

We found that pseudo-relevance feedback (PRF or blind-query expansion) could be a useful technique for enhancing retrieval effectiveness. In this study, we adopted Rocchio’s approach [11] 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 models and enlarged the query by the 20 to 120 terms extracted from the 5 to 10 best-ranked articles (see Table 3).

For the Hungarian collection, percentage improvement varied from +7.6% (IneC2 model, 0.3897 vs. 0.4193) to +10.5% (Okapi model, 0.3629 vs. 0.4010). For the Bulgarian corpus, enhancement increased from +18% (LM model, 0.3368 vs. 0.4004) to +21.7% (LM model, 0.3368 vs. 0.4098). For the Czech language, the variation percentages ranged from 6.0% (Okapi model, 0.3355 vs. 0.3557) to +10.3% (0.3355 vs. 0.3702). As shown in Table 3, the performance differences before and after query expansion were always statistically significant.

5 Data Fusion and Official Results

It is usually assumed that combining result lists computed by different search models (data fusion) should improve retrieval effectiveness, for three reasons [12]. This first is a skimming process, in which only the m top-ranked items retrieved from each ranked list are considered. In this case, we would combine the best answers obtained from various document representations. The second is the chorus effect, by which different retrieval schemes would retrieve the same item, and as such provide stronger evidence that the corresponding document is indeed relevant. The third is an opposite or dark horse effect, which may also play a role. A given retrieval model may provide unusually high and accurate estimates of a document’s relevance. Thus, a combined system could possibly return more pertinent items by accounting for documents obtaining a relatively high score.

To present the official runs described in Table 4 we combined three probabilistic models, representing both the parametric (Okapi and DFR) and non-parametric (LM) probabilistic approaches. All runs were fully automated and in

Table 4. Description and MAP of Our Best Official Monolingual Runs

Language	Index	Query	Model	Query exp.	MAP	comb. MAP
Hungarian UniNEhu2	dec.	TD	LM	5 docs/70 terms	0.4315	Z-score
	word	TD	GL2	5 docs/100 terms	0.4376	0.4716
	4-gram	TD	Okapi	3 docs/120 terms	0.4233	
Bulgarian UniNEbg1	4-gram	TD	Okapi	3 docs/150 terms	0.3169	Z-score
	word	TD	PB2	5 docs/60 terms	0.3750	0.4128
	word	TD	LM	10 docs/50 terms	0.4098	
Czech UniNEcz3	word	TD	LM	5 docs/20 terms	0.4070	Z-score
	4-gram	TD	Okapi	5 docs/70 terms	0.3672	0.4225
	word	TD	GL2	5 docs/50 terms	0.4085	

all cases applied the same data fusion approach (Z-score [13]). For the Hungarian corpus however we occasionally applied our decompounding approach (denoted by “dec” in the “Index” column). As shown in Table 4, for a data fusion strategy retrieval performance is clearly better for the Hungarian language, moderate for the Bulgarian and only slightly better for the Czech language.

6 Conclusion

In this eighth CLEF evaluation campaign we analyze various probabilistic IR models using three different test-collections written in three East European languages (Hungarian, Bulgarian and Czech). We suggest a new stemmer for the Bulgarian language that removes some very frequently appearing derivational suffixes. For the Czech language, we design and implement two different stemmers.

Our various experiments demonstrate that the IneC2 model derived from *Divergence from Randomness* (DFR) paradigm tends to produce the best overall retrieval performances (see Tables 1 or 2). The statistical language model (LM) used in our experiments usually provides inferior retrieval performance to that obtained with the Okapi or DFR approach.

For the Bulgarian language (Table 1), our new and more aggressive stemmer tends to produce better MAP compared to a light stemming approach (around +6% in relative difference). For the Hungarian language (Table 1), applying an automated decompounding procedure improves the MAP around +10.8% when compared to a word-based approach. For the Czech language however performance differences between a light and a more aggressive stemmer removing both inflectional and some derivational suffixes are rather small (Table 2). Moreover, performance differences are also small when compared to those achieved with a 4-gram approach. The pseudo-relevance feedback may improve the MAP, depending on the parameter settings used (Table 3).

Acknowledgments. This research was supported in part by the Swiss National Science Foundation under Grant #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., Abdou, S.: Experiments with Monolingual, Bilingual, and Robust Retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 137–144. Springer, Heidelberg (2007)
3. 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)
4. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview. In: Peters, C., et al. (eds.) *CLEF 2007*. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)
5. Savoy, J.: Searching Strategies for the Hungarian Language. *Information Processing & Management* 44, 310–324 (2008)
6. Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a Way of Life: Okapi at TREC. *Information Processing & Management* 36, 95–108 (2002)
7. 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)
8. Hiemstra, D.: Using Language Models for Information Retrieval. PhD Thesis (2000)
9. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management* 33, 495–512 (1997)
10. McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. *IR Journal* 7, 73–97 (2004)
11. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: *Proceedings TREC-4*, Gaithersburg, pp. 25–48 (1996)
12. Vogt, C.C., Cottrell, G.W.: Fusion via a Linear Combination of Scores. *IR Journal* 1, 151–173 (1999)
13. 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. (eds.) *CLEF 2005*. LNCS, vol. 4022, pp. 131–140. Springer, Heidelberg (2006)

Applying Query Expansion Techniques to 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 the Monolingual tasks at CLEF 2007. We submitted results for the following languages: Hungarian, Bulgarian and Czech. We focused on studying different query expansion techniques: Probabilistic Relevance Feedback (PRF) and Mutual Information Relevance Feedback (MI-RF) to improve retrieval performance. After an analysis of our experiments and of the official results at CLEF 2007, we achieved considerably improved scores by using query expansion techniques for different languages (Hungarian, Bulgarian and Czech).

1 Introduction

Query expansion (QE) is a technique commonly used in Information Retrieval (IR) to improve retrieval performance by reformulating the original query adding new terms or re-weighting the original terms. Query expansion terms can be automatically extracted from documents or taken from knowledge resources.

In our seventh participation at CLEF, we focused on comparing two different query expansion strategies: Probabilistic Relevance Feedback (PRF) and Mutual Information Relevance Feedback (MI-RF). Specifically, we participated in tasks for the following languages: Hungarian, Bulgarian and Czech.

We used the IR-n system [2]. It is a Passage Retrieval (PR) system which uses passages with a fixed number of sentences. This provides the passages with some syntactical content.

This paper is organized as follows: the next section describes the task developed by our system and the training carried out for CLEF 2007. The results obtained are then presented. Finally, we present the conclusions and the future work.

2 Relevance Feedback

Query expansion techniques such as Relevance Feedback (RF) can substantially improve retrieval effectiveness. Most of the IR systems commonly implemented query expansion techniques. RF is usually performed in the following way:

- A search using the original query is performed, selecting the n terms from top-ranked documents.
- The n terms are added to the original query to formulate a new query.
- The new query is performed to produce a new ranked list of documents.

An important factor is how to assign the weight to the selected terms with respect to the terms from the initial query. In this work, we compare two formulas in order to calculate this weight (w_t): Probabilistic and Mutual Information.

2.1 Probabilistic Relevance Feedback (PRF)

This is the term relevance weighting formula proposed by Robertson and Sparck Jones in [3]. The relevance weight of term t is given by:

$$w_t = \frac{(m_t + 0.5) \cdot (n - n_t - m + m_t + 0.5)}{(m - m_t + 0.5) \cdot (n_t - m_t + 0.5)} \quad (1)$$

where n is the number of documents in the collection, m is the number of documents considered as relevants (in this case 10 documents), n_t is the number of documents which the term t appears and m_t is the number of relevant documents in which the term t appears. w_t will be better for those terms which have a higher frequency in the relevant documents than the whole collection.

2.2 Mutual Information Relevance Feedback (MI-RF)

This is based on the idea the co-occurrence between two terms can determine the semantic relation that exists between them [1]. The mutual information score grows with the increase in frequency of word co-occurrence. If two words co-occur mainly due to chance their mutual information score will be close to zero. If they occur predominantly individually, then their mutual information will be a negative number. The standard formula for calculating mutual information is:

$$MI(x, y) = \log\left(\frac{P(x, y)}{P(x) \cdot P(y)}\right) \quad (2)$$

where $P(x, y)$ is the probability that words x and y occur together; $P(x)$ and $P(y)$ are the probabilities that x and y occur individually. The relevance weight w_t of each term t is calculated adding the MI between t and each term of the query.

3 Experiments

The aim of the experimental phase was to set up the optimum value of the input parameters for each collection. CLEF 2005 and 2006 (Hungarian and Bulgarian) collections were used for training. Query expansion techniques were also evaluated 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 dfr weighting model. It has two parameters: c and $avgld$.
- **Query expansion parameters**: If **exp** has value 1, this denotes we use PRF based on passages. If **exp** has value 2, the PRF is based on documents. And, if **exp** has value 3, MI-RF query expansion is used. Moreover, **np** and **nd** denote the k terms (nd) extracted from the best ranked passages or documents (np) from the original query.
- **Evaluation measure**: Mean average precision (**avgP**) is the evaluation measure used in order to evaluate the experiments.

Table 1. Highest AvgP obtained with training data CLEF 2006

language	sp	wm	C	avgld	exp	np	nd	avgP
Hungarian	8	dfr	2	300				0.3182
Hungarian	8	dfr	2	300	2	10	10	0.3602
Hungarian	8	dfr	2	300	3	10	10	0.3607
Bulgarian	30	dfr	1.5	300				0.1977
Bulgarian	30	dfr	1.5	300	2	10	10	0.2112
Bulgarian	30	dfr	1.5	300	3	10	10	0.2179

Table 1 shows the best configuration for each language. The best weighting scheme for Hungarian and Bulgarian was dfr. For Hungarian, we used 8 as passage size. For Bulgarian, we set up 30 as passage size. Finally, the configuration used for Czech was the same as for Hungarian (dfr as weighting scheme and 30 as passage size).

4 Results at CLEF 2007

We submitted four runs for each language in our participation at CLEF 2007. The best parameters, i.e. those that gave the highest mean AvgP score in system training, were used in all cases. The name of the runs has this pattern: IRnxyyyN. xx is the language (BU, HU or CZ), yyyy is the query expansion (*nexp*: not used, *exp2*: PRF, *exp3*: MI-RF) and N means the tag *narrative* was used.

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 baseline in all languages. The best percentage of improvement in AvgP is 40.09% for Hungarian.

Table 2. CLEF 2007 official results. Monolingual tasks.

Language Run		AvgP Dif	
Hungarian	IRnHUnexp (baseline)	33.90	
	IRnHUexp2	38.94	+14.88%
	IRnHUexp3	39.42	+16.29%
	IRnHUexp2N	40.09	+18.26%
Bulgarian	IRnBUnexp (baseline)	21.19	
	IRnBUexp2	25.97	+22.57%
	IRnBUexp3	26.35	+24.36%
	IRnBUexp2N	29.81	+40.09%
Czech	IRnCZnexp (baseline)	20.92	
	IRnCZexp2	24.81	+18.61%
	IRnCZexp3	24.84	+18.76%
	IRnCZexp2N	27.68	+32.36%

5 Conclusions and Future Work

In this eighth CLEF evaluation campaign, we compared different query expansion techniques in our system for Hungarian, Bulgarian and Czech (see Table 1). Specifically, we compare two query expansion techniques: Probabilistic Relevance Feedback (PRF) and Mutual Information Relevance Feedback (MI-RF).

The results of this evaluation indicate that for the Hungarian, Bulgarian and Czech our approach proved to be effective (see Table 2) because the results are above baseline. For all languages, the increases in mean AvgP from both query expansion methods were about the same.

In the future we intend to test this approach in other languages such as 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 Spanish Government, project TEXT-MESS (TIN-2006-15265-C06-01) and by the Valencia Government under project number GV06-161.

References

1. Gale, W.A., Church, K.W.: Identifying word correspondence in parallel texts. In: HLT 1991: Proceedings of the workshop on Speech and Natural Language, Morristown, NJ, USA, pp. 152–157. ACL (1991)
2. Llopis, F.: IR-n: Un Sistema de Recuperación de Información Basado en Pasajes. PhD thesis, University of Alicante (2003)
3. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms, pp. 143–160. Taylor Graham Publishing, London (1988)

Bulgarian, Hungarian and Czech Stemming Using YASS

Prasenjit Majumder, Mandar Mitra, and Dipasree Pal

CVPR Unit, Indian Statistical Institute, Kolkata
{`prasenjit_t,mandar,dipasree_t`}@`isical.ac.in`

Abstract. This is the second year in a row we are participating in CLEF. Our aim is to test the performance of a statistical stemmer on various languages. For CLEF 2006, we tried the stemmer on French [1]; while for CLEF 2007, we did experiments for the Hungarian, Bulgarian and Czech monolingual tasks. We find that, for all languages, YASS produces significant improvements over the baseline (unstemmed) runs. The performance of YASS is also found to be comparable to that of other available stemmers for all the three east European Languages.

1 Introduction

Stemming is arguably a recall enhancing device in text retrieval. Most commonly used stemmers are rule-based and therefore language specific. Such stemmers are unlikely to be available for resource poor languages. In earlier work, therefore, we proposed YASS [2], a statistical stemmer. As YASS does not assume any language specific information, we expect the approach to work for multiple languages. The motivation behind our experiments at CLEF 2006 last year was to test this hypothesis. Since our hypothesis was supported by last year's experiments, this year, for CLEF 2007, we planned on monolingual retrieval for more languages which we know nothing about.

The main stumbling block in our experiments was the encoding issue. We modified our systems to work with UTF-8 data. During the official submission, we could not complete the Bulgarian runs and submitted only six official runs for Hungarian and Czech. After the relevance judgements were released, we tuned the statistical stemmer for each of the three languages.

Three retrieval models were used in our study, viz. BM25, DFR-In_expC2, and TF.IDF (Lnu.ltn). Our experiments were conducted using the SMART[3] system for the tf.idf model, and the Terrier[4] system for the rest of the models.

We give a brief overview of YASS in the next section. Section 3 presents and analyses the results of all the runs (both official and unofficial) for the three languages. Section 4 concludes the paper.

2 YASS

YASS (Yet Another Suffix Stripper) [2] is a statistical stemmer that is based on a string distance measure. Using this measure, YASS clusters a lexicon created

from a text corpus. Each cluster is expected to contain all the morphological variations of a root word. The clustering method (agglomerative hierarchical clustering) requires a threshold value (referred to as θ henceforth) as a parameter. If training data is available, this parameter may be tuned to improve the performance of the stemmer. The following subsections will describe the string distance used and the training procedure for threshold selection.

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, $D = \frac{6}{8} \times (\frac{1}{2^0} + \dots + \frac{1}{2^{13-8}}) = 1.4766$.

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

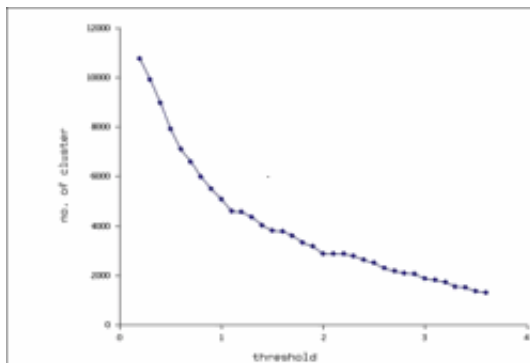


Fig. 1. Number of clusters at various thresholds for Hungarian

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 the literature, namely, *single-linkage*, *average-linkage*, and *complete-linkage*. We choose the complete-linkage algorithm for our experiments.

2.3 Training

We have mentioned earlier that YASS needs no linguistic input as it is a statistical stemmer. However, before running YASS on a new language, we need to train it for getting a suitable clustering threshold. As training data was not available for Bulgarian and Czech before the official submission, we set the threshold value to 1.5 based on our earlier experience with English and French. For Hungarian, we used the CLEF2006 data for training. A lexicon was extracted from the corpus and clustered using various thresholds, resulting in a set of stemmers. A suitable threshold was chosen based on the performance of these stemmers. After the relevance judgements were released for all the languages, we tuned the threshold of YASS for Bulgarian and Czech as well, using this year’s data.

Hungarian. The same Hungarian corpus is used for the 2005, 2006, and 2007 tasks. The lexicon extracted from the corpus has 536,678 surface words. The lexicon was clustered using various threshold settings, and the number of clusters versus threshold curve is shown in Figure 1. The step-like regions around 0.8, 1.1, 1.5, 2.0 suggest that the number of clusters is stable around these threshold values. These values may thus be chosen as candidate thresholds for clustering. After clustering the lexicon using these four threshold values, the lexicon size

Table 1. Threshold vs. MAP for Hungarian

Threshold	Mean AvgP (MAP)
0.8	0.2692
1.1	0.2835
1.5	0.2777
2	0.2735

Table 2. Hungarian runs for training on CLEF 2005 dataset

YASS			
Run Name	MAP	R-prec	% Rel_Ret
noStem(T+D+N)	0.2472	0.2531	72.8434
$\theta = 0.8$ (T+D+N)	0.3211	0.3231	83.8125
$\theta = 1.1$ (T+D+N)	0.3246	0.3247	86.0489
$\theta = 1.5$ (T+D+N)	0.3179	0.3190	86.6879
$\theta = 2.0$ (T+D+N)	0.3005	0.3068	83.8125
noStem(T+D)	0.2170	0.2285	69.1160
$\theta = 0.8$ (T+D)	0.3121	0.3162	81.0436
$\theta = 1.1$ (T+D)	0.3241	0.3270	84.2385
$\theta = 1.5$ (T+D)	0.3268	0.3309	85.6230
$\theta = 2.0$ (T+D)	0.3048	0.3074	84.1320

TORDAI stemmer			
Run Name	MAP	R-prec	% Relevant Docs Retrieved
Heavy minus hyphen	0.3099	0.3048	83.1
4-Gram	0.3303	0.338	83.6
5-Gram	0.3002	0.3057	82.4

gets reduced to 225489, 169619, 130278, and 76782 classes respectively. The stemmers thus prepared are used in four different official runs.

The official topics for the Hungarian monolingual run at CLEF-2006 were topic numbers 301 to 325 and 351 to 375. Table 1 suggests that the performance of YASS does not change much as the threshold varies between 1.1 and 2.

We also tested these stemmers on CLEF queries 251 to 300. These queries were used in the CLEF 2005 monolingual Hungarian task. Table 2 gives the results of these runs, as well as the best results reported by Tordai et al. [5] for the same task at CLEF 2005. This table also suggests that setting $\theta = 1.1$ or 1.5 would be appropriate for Hungarian.

3 Experiments

This section describes all the runs we performed for all the three languages. Besides the six official runs, we performed several other experiments using the three east European languages, to better understand the performance of YASS.

Table 3. Official results on 2007 CLEF data

Hungarian Runs Submitted (<i>nnn.ltn</i>)			
Run Name	MAP	R-prec	% Rel_Ret
ISI.YASSHUN	0.1712	0.1974	72.22
ISI.YASSTDHUN	0.1695	0.1943	72.88
ISI.ISIDWLDHSTEMGZ	0.1605	0.1858	66.84
Runs Submitted (<i>Lnu.ltn</i>)			
ISI.CZTD [YASS] (T+D)	0.3224	0.3102	87.13
ISI.ISICL [dnlded] (T+D+N)	0.3362	0.3326	89.37
ISI.ISICZNS [nostem] (T+D+N)	0.2473	0.2540	76.64

Table 4. Word stems generated by YASS

Hungarian		Czech	
politikusokról, politikai	politi	Kosteličových, Kosteličovi	Kostelič
atomhulladékot	atomhulladék	prezidenští, prezidenta, prezidentského	preziden
megszűnése	megsz	kandidáti, kandidáta	kandidát
elnökjelöltek, elnökjelölt	elnökjelöl	vesmírní, vesmírných, vesmíru	vesmír
királynő, királyságbeli	kir / király	turistech, turisté	turist

3.1 Official Runs

In the first Hungarian run *ISI.YASSTDHUN*, we indexed only the <title> and <desc> fields of the queries. For the second run, *ISI.YASSHUN* we indexed the <title>, <desc>, and <narr> fields of the queries. In both cases the clustering threshold was set to 1.5. For the third run, *ISI.ISIDWLDHSTEMGZ*, we made use of a Hungarian stemmer available from the web¹.

The Czech runs are analogous: the first run uses only the <title> and <desc> fields; the second and third runs use the complete query. The second run makes use of an existing stemmer² instead of YASS. The final run was a baseline run where no stemming was used.

Table 3 shows the results of our official runs. These results confirm our hypothesis that YASS will work for a variety of languages, provided the languages are primarily suffixing in nature. Table 4 provides some examples of words and their roots obtained using YASS. These words were selected from queries on which the stemmed run significantly outperformed the unstemmed run.

3.2 Other Runs

Other groups that have reported results for these three east European languages in this volume include [6], [7], and [8]. We were particularly interested in the

¹ <http://snowball.tartarus.org/algorithms/hungarian/stemmer.html>

² <http://members.unine.ch/jacques.savoy/clef/index.html>

Table 5. Performance of YASS for various models and parameter settings

Hungarian				
Models	topics	1.5	2.0	no-stem
DFR	TD	0.3358	0.3535	0.2461
	TDN	0.3728	0.3902	0.2813
OKAPI	TD	0.2920	0.3138	0.1992
	TDN	0.3274	0.3445	0.2285
TFIDF	TDN	0.3600	0.3638	0.2647
Czech				
Models	topics	1.5	2.0	no-stem
DFR	TD	0.3337	0.3483	0.2320
	TDN	0.3574	0.3674	0.2525
OKAPI	TD	0.3199	0.3306	0.2162
	TDN	0.3332	0.3464	0.2454
TFIDF	TDN	0.3390	0.3381	0.2473
Bulgarian				
Models	topics	1.5	2.0	no-stem
DFR	TD	0.3533	0.3526	0.2586
	TDN	0.3626	0.3649	0.2862
OKAPI	TD	0.3289	0.3330	0.2346
	TDN	0.3439	0.3465	0.2594

results reported by Dolamic and Savoy [6] for two reasons. First, their work motivated us to explore retrieval models besides the traditional tf.idf method implemented in the SMART system. Secondly, they present results obtained using linguistically-based stemming / decompounding algorithms. It would be interesting to compare the performance of these methods with that of a purely statistical method such as YASS.

Accordingly, after the relevance judgments for the data sets were distributed, we performed some additional experiments for the three languages. The primary aim of these experiments was two-fold: (i) To use YASS with alternative retrieval approaches, specifically the BM25 weighting method, and the Divergence from Randomness (DFR) model. (ii) To compare YASS with the stemming / decompounding methods described by Dolamic and Savoy.

For these experiments, we used the BM25 scheme and a variant of the Divergence from Randomness model (DFR-In_expC2) as implemented in the Terrier-2.0 system. The c parameter of DFR-In_expC2 was set to the Terrier default value 1.0 for most runs (see Table 5); however, when comparing results with those reported by Dolamic and Savoy, we used $c = 1.5$ as this was the c value used in their work (see Table 6).

Besides exploring alternative retrieval strategies, we also tried a range of clustering thresholds. The results for $\theta = 1.5$ and $\theta = 2.0$ are reported in Table 5. These experiments suggest that 2.0 is a good choice of the parameter θ in YASS for the three east European languages, irrespective of retrieval models.

3.3 Comparing and Analysis of Results

With the clustering threshold θ set to 2.0, we compared YASS with the stemmers described in [6]. We chose the 12 best runs from that paper for comparison (the 4-gram based runs are not considered, since this approach was less effective than the other approaches). All these runs are based on the DFR model, since it yielded the best performance reported in [6]. As mentioned above, we use $c = 1.5$ for these runs (as suggested in [6]); however, the mean document length (mean dl) parameter is unchanged from the default setting in Terrier (this parameter was set to 213, 135, 152 for Czech, Bulgarian and Hungarian, resp., in [6]).

Table 6 compares the results obtained using YASS with those reported by Dolamic and Savoy. The performance differences were found to be statistically insignificant (based on a t -test) for the four Czech and Bulgarian runs.

Of the best four Hungarian runs reported in [6], two runs (TD, TDN) use a stemmer [9,10], and two runs (TD, TDN) use a de-compounding algorithm [11]. Once again, no significant difference was found between these methods and YASS when only the title and description fields of the query were indexed (runs labeled TD). However, the decompounding run using the full query (TDN) was found to be significantly better than YASS. A more detailed analysis of this difference

Table 6. Comparison between YASS and Dolamic et al.

Bulgarian runs				
Model	topics	light/word	deriv./word	YASS
DFR	TD	0.3423	0.3606	0.3613
	TDN	0.3696	0.3862	0.3748
Czech runs				
Model	topics	light	derivational	YASS
DFR	TD	0.3437	0.3342	0.3523
	TDN	0.3678	0.3678	0.3702
Hungarian runs				
Model	topics	stemmer(word)	de-compound	YASS
DFR	TD	0.3525	0.3897	0.3588
	TDN	0.4031	0.4271	0.3951

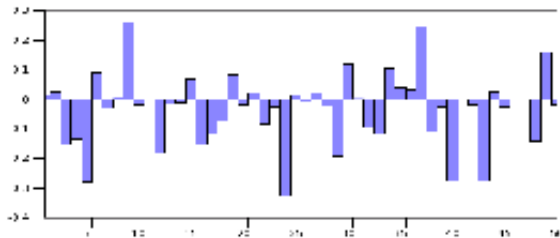


Fig. 2. Difference in AvgP for individual queries for YASS and de-compounding

is presented in Figure 2, which shows that, out of 50 queries, YASS performed better in 21 cases, while the decompounding method did better in 29 cases.

4 Conclusion

Overall, we found that YASS performs as well as any linguistic stemmers for the three east European languages viz. Hungarian, Bulgarian and Czech. Our explorations of alternative retrieval approaches (besides the traditional tf.idf method) yielded promising results. In future work, we hope to undertake a more complete investigation of YASS within the context of these models.

References

1. Majumder, P., Mitra, M., Datta, K.: Statistical vs. rule-based stemming for monolingual french retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 107–110. Springer, Heidelberg (2007)
2. Majumder, P., Mitra, M., Parui, S.K., Kole, G., Mitra, P., Datta, K.: Yass: Yet another suffix stripper. *ACM Trans. Inf. Syst.* 25(4), 18 (2007)
3. Salton, G. (ed.): *The SMART Retrieval System—Experiments in Automatic Document Retrieval*. Prentice Hall Inc., Englewood Cliffs (1971)
4. Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research directions in terrier. In: Baeza-Yates, R., et al. (eds.) *Novatica/UPGRADE Special Issue on Web Information Access (Invited Paper)* (2007)
5. Tordai, A., de Rijke, M.: Four Stemmers and a Funeral: Stemming in Hungarian 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. 179–186. Springer, Heidelberg (2006)
6. Dolamic, L., Savoy, J.: Stemming approaches for east european languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 37–44. Springer, Heidelberg (2008)
7. Ircing, P., Muller, L.: Czech Monolingual Information Retrieval Using Off-The-Shelf Components - the University of West Bohemia at CLEF 2007 Ad-Hoc track. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
8. Ceska, P., Pecina, P.: Charles University at CLEF 2007 Ad-Hoc Track. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
9. Savoy, J., Abdou, S.: Experiments with monolingual, bilingual, and robust retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 137–144. Springer, Heidelberg (2007)
10. Savoy, J.: Searching strategies for the hungarian language. *Inf. Process Manage* 44(1), 310–324 (2008)
11. 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)

Sampling Precision to Depth 10000 at CLEF 2007

Stephen Tomlinson

Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
<http://www.opentext.com/>

Abstract. We conducted an experiment to test the completeness of the relevance judgments for the monolingual Bulgarian, Czech and Hungarian information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2007. In the ad hoc retrieval tasks, the system was given 50 natural language queries, and the goal was to find all of the relevant documents (with high precision) in a particular document set. For each language, we submitted a sample of the first 10000 retrieved items to investigate the frequency of relevant items at deeper ranks than the official judging depth (of 60 for Czech and 80 for Bulgarian and Hungarian). The results suggest that, on average, the percentage of relevant items assessed was less than 60% for Czech, 70% for Bulgarian and 85% for Hungarian. These levels of completeness are in line with the estimates that have been made for some past test collections which are still considered useful and fair for comparing retrieval methods.

1 Introduction

Livelihood ECM - eDOCS SearchServerTM is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other components of the Livelihood ECM - eDOCS Suite¹.

SearchServer works in Unicode internally [4] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (CLEF [1], NTCIR [5] and TREC [7]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This paper describes an experiment conducted with SearchServer for testing the completeness of the relevance judgments for the monolingual Bulgarian, Czech and Hungarian information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2007.

¹ Livelihood, Open TextTM and SearchServerTM are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

2 Methodology

2.1 Data

The CLEF 2007 Ad-Hoc Track document sets consisted of tagged (SGML-formatted) news articles in 3 different languages: Bulgarian, Czech and Hungarian. Table 1 gives the sizes.

Table 1. Sizes of CLEF 2007 Ad-Hoc Track Test Collections

Language	Text Size (uncompressed)	Documents	Topics	Rel/Topic
Bulgarian	265,368,055 bytes	87,281	50	20 (lo 2, hi 62)
Czech	151,914,429 bytes	81,735	50	15 (lo 2, hi 47)
Hungarian	106,631,823 bytes	49,530	50	18 (lo 1, hi 66)

The CLEF organizers created 50 natural language “topics” (numbered 401-450) and translated them into many languages. Several research groups submitted their top-1000 ranked results for each topic in June 2007. The organizers pooled some of the top-ranked documents and judged them as relevant or not relevant. Table 1 gives the official average number of relevant documents for each language (along with the lowest and highest number of relevant documents of any topic). For more information on the CLEF ad hoc test collections, please see the track overview paper [2].

2.2 Base Run

We wanted our base run for each language to be as strong performing as possible. We conducted various diagnostic experiments on past CLEF collections for Bulgarian and Hungarian (a continuation of the experiments described in [8]) which suggested that our best option was to have the base runs be a fusion of a stemming run and a 4-gram run.

For the stemming runs, we used all 3 topic fields (Title, Description and Narrative). We thank Jacques Savoy for providing experimental algorithmic stemmers and stopword lists [6] for all 3 languages. Note that for Czech, our port of the stemmer was accent-insensitive.

For the 4-gram runs, we just used the Title and Description fields (the diagnostic experiments had found that the Narrative field reduced the average precision of the 4-gram technique). For 4-grams, a different index was used which primarily consisted of the 4-grams of terms, e.g. the word ‘search’ would produce index terms of ‘sear’, ‘earc’ and ‘arch’. No stopwords were applied. For Bulgarian, we did not index the breve accent for the 4-gram run (unlike for our stemming run).

To implement fusion, for each language we retrieved the top-10000 ranked documents for both the stemming run and 4-gram run and added together the rsv (retrieval status value) scores. The resulting top-10000 documents were used for our base run for each language.

2.3 Sample Run

For each language, we created a sample run whose first 100 rows contained the following rows of the base run for the language in the following order:

1, 2, ..., 10,
 20, 30, ..., 100,
 200, 300, ..., 1000,
 2000, 3000, ..., 10000,
 15, 25, ..., 95,
 150, 250, ..., 950,
 1500, 2500, ..., 9500,
 125, 175, ..., 975,
 1250, 1750, ..., 9750.

The remainder of the sample run was padded with the top-ranked remaining rows from the base run until 1000 rows had been retrieved (i.e. rows 11, 12, 13, 14, 16, ..., 962 of the base run).

This ordering (e.g. the placement of the sample from depth 10000 before the sample from depth 15) was chosen because of uncertainty of how deep the judging would be. As long as the top-37 were judged, we would have sampling to depth 10000. The extra sample points would just improve the accuracy.

Our sample run for each language was submitted to the CLEF organizers for assessing in June 2007. (Our base run for each language was not submitted.)

3 Results

When we received the relevance judgments in August 2007 we checked the judging depth of our sample runs. We found that the top-60 rows were judged for each topic for Czech and the top-80 rows were judged for each topic for Bulgarian and Hungarian.

Tables 2, 3 and 4 show the results of the sampling for each language. The columns are as follows:

- “Depth Range”: The range of depths being sampled. The 11 depth ranges covered from 1 to 10000.
- “Samples”: The depths of the sample points from the depth range. The samples were always uniformly spaced. They always ended at the last point of the depth range. The total number of sample points (over the 11 rows of the table) adds to 60 for Czech and 80 for Bulgarian and Hungarian.
- “# Rel”: The number of each type of item retrieved from the sample points over the 50 topics. The item type codes are R (relevant), N (non-relevant) and U (unjudged, of which there were always 0). The sum of the item type counts is always 50 times the number of sample points for the depth range (because there were 50 topics for each language).
- “Precision”: Estimated precision of the depth range ($R/(R+N+U)$).

Table 2. Marginal Precision of Bulgarian Base Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	107R, 143N, 0U	0.428	1	2.1
6-10	6, 7, ..., 10	92R, 158N, 0U	0.368	1	1.8
11-50	15, 20, ..., 50	70R, 330N, 0U	0.175	5	7.0
51-100	55, 60, ..., 100	28R, 472N, 0U	0.056	5	2.8
101-200	125, 150, ..., 200	5R, 195N, 0U	0.025	25	2.5
201-500	225, 250, ..., 500	2R, 598N, 0U	0.003	25	1.0
501-900	525, 550, ..., 900	2R, 798N, 0U	0.003	25	1.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	1R, 199N, 0U	0.005	500	10.0
3001-6000	3500, 4000, ..., 6000	0R, 300N, 0U	0.000	500	0.0
6001-10000	6500, 7000, ..., 10000	0R, 400N, 0U	0.000	500	0.0

- “Wgt”: The weight of each sample point. The weight is equal to the difference in ranks between sample points. Each sample point can be thought of as representing this number of rows, and in particular, the rows consisting of the sample point itself plus the preceding unsampled rows. The weights are higher in some cases for Czech than for Bulgarian and Hungarian because we had fewer sample points for Czech (60 instead of 80).
- “EstRel/Topic”: Estimated number of relevant items retrieved per topic for this depth range. This is the Precision multiplied by the size of the depth range. Or equivalently, it is $(R * Wgt) / 50$.

Because each sample point was at the deep end of the range of rows it represented, the sampling should tend to underestimate precision for each depth range (assuming that precision tends to fall with depth, which appears to have been the case for all 3 languages). Hence our estimates of the number of relevant items in the original base run should tend to be on the low side.

Table 5 compares the estimated number of relevant items in the base run to the official number of relevant items for each language. The first row, “Estimated Rel@10000”, shows the sums of the estimated number of relevant items per topic over all depth ranges. The second row, “Official Rel/Topic”, shows the official number of relevant items per topic. The final row, “Percentage Judged”, just divides the official number of relevant items by the estimated number in the first 10000 retrieved (e.g. for Bulgarian, $20.2/29.3=69\%$). This number should tend to be an overestimate of the percentage of all relevant items that are judged (on average per topic) for two reasons: there may be relevant items that were not matched by our base run in the first 10000 rows, and (as previously mentioned) our Estimated Rel@10000 already tended to be on the low side.

3.1 Remarks

These estimates of the judging coverage (i.e. percentage of relevant items assessed) for the CLEF 2007 collections (55% for Czech, 69% for Bulgarian, 83%

Table 3. Marginal Precision of Czech Base Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	110R, 140N, 0U	0.440	1	2.2
6-10	6, 7, ..., 10	71R, 179N, 0U	0.284	1	1.4
11-50	15, 20, ..., 50	48R, 352N, 0U	0.120	5	4.8
51-100	55, 60, ..., 100	10R, 490N, 0U	0.020	5	1.0
101-200	150, 200	3R, 97N, 0U	0.030	50	3.0
201-500	250, 300, ..., 500	1R, 299N, 0U	0.003	50	1.0
501-900	550, 600, ..., 900	3R, 397N, 0U	0.007	50	3.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	0R, 200N, 0U	0.000	500	0.0
3001-6000	3500, 4000, ..., 6000	1R, 299N, 0U	0.003	500	10.0
6001-10000	7000, 8000, ..., 10000	0R, 200N, 0U	0.000	1000	0.0

Table 4. Marginal Precision of Hungarian Base Run at Various Depths

Depth Range	Samples	# Rel	Precision	Wgt	EstRel/Topic
1-5	1, 2, ..., 5	133R, 117N, 0U	0.532	1	2.7
6-10	6, 7, ..., 10	89R, 161N, 0U	0.356	1	1.8
11-50	15, 20, ..., 50	55R, 345N, 0U	0.138	5	5.5
51-100	55, 60, ..., 100	25R, 475N, 0U	0.050	5	2.5
101-200	125, 150, ..., 200	3R, 197N, 0U	0.015	25	1.5
201-500	225, 250, ..., 500	12R, 588N, 0U	0.020	25	6.0
501-900	525, 550, ..., 900	2R, 798N, 0U	0.003	25	1.0
901-1000	950, 1000	1R, 99N, 0U	0.010	50	1.0
1001-3000	1500, 2000, ..., 3000	0R, 200N, 0U	0.000	500	0.0
3001-6000	3500, 4000, ..., 6000	0R, 300N, 0U	0.000	500	0.0
6001-10000	6500, 7000, ..., 10000	0R, 400N, 0U	0.000	500	0.0

Table 5. Estimated Percentage of Relevant Items that are Judged, Per Topic

	Bulgarian	Czech	Hungarian
Estimated Rel@10000	29.3	27.4	21.9
Official Rel/Topic	20.2	15.2	18.2
Percentage Judged	69%	55%	83%

for Hungarian) are much higher than the estimates we produced for the TREC 2006 Legal and Terabyte collections using a similar approach (18% for TREC Legal and 36% for TREC Terabyte) [9]. They are similar to the estimates we produced for the NTCIR-6 collections (58% for Chinese, 78% for Japanese, 100% for Korean) [10].

These estimates of the judging coverage for the CLEF 2007 collections are also similar to what [11] estimated (using a different approach) for depth-100

pooling on the old TREC collections of approximately 500,000 documents: “it is likely that at best 50%-70% of the relevant documents have been found; most of these unjudged relevant documents are for the 10 or so queries that already have the most known answers.”

Fortunately, [11] also found for such test collections that “overall they do indeed lead to reliable results.” [3] also considers the “levels of completeness” in some older TREC collections to be “quite acceptable” even though additional judging found additional relevant documents. And we can confirm that we have gained a lot of insights from the CLEF test collections over the years, particularly when conducting topic analyses such as described in [8].

3.2 Error Analysis

We should note that our sampling was very coarse at the deeper ranks, e.g. for Czech, 1 relevant item out of 300 samples in the 3001-6000 range led to an estimate of 10 relevant items per topic in this range. If the sampling had turned up 0 or 2 relevant items, a minor difference, the estimate would have been 0 or 20 relevant items per topic in this range, leading to a substantially different sum (17.4 or 37.4 instead of 27.4). We leave the computation of confidence intervals for our estimates, along with analysis of the variance across topics, as future work.

4 Conclusions

We conducted an experiment to test the completeness of the relevance judgments for the monolingual Bulgarian, Czech and Hungarian information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2007. For each language, we submitted a sample of the first 10000 retrieved items to investigate the frequency of relevant items at deeper ranks than the official judging depth (of 60 for Czech and 80 for Bulgarian and Hungarian). Based on the results, we estimated that the percentage of relevant items assessed was less than 55% for Czech, 69% for Bulgarian and 83% for Hungarian. These levels of completeness are in line with the estimates that have been made for some past test collections which are still considered useful and fair for comparing retrieval methods.

References

1. Cross-Language Evaluation Forum web site, <http://www.clef-campaign.org/>
2. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)
3. Harman, D.K.: The TREC Test Collections. In TREC: Experiment and Evaluation in Information Retrieval (2005)
4. Hodgson, A.: Converting the Fulcrum Search Engine to Unicode. In: Sixteenth International Unicode Conference (2000)

5. NTCIR (NII-NACSIS Test Collection for IR Systems),
<http://research.nii.ac.jp/~ntcadm/index-en.html>
6. Savoy, J.: CLEF and Multilingual information retrieval resource page,
<http://www.unine.ch/info/clef/>
7. Text REtrieval Conference (TREC), <http://trec.nist.gov/>
8. Tomlinson, S.: Bulgarian and Hungarian Experiments with Hummingbird SearchServerTM 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)
9. Tomlinson, S.: Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track. In: Proceedings of TREC 2006 (2006)
10. Tomlinson, S.: Sampling Precision to Depth 9000: Evaluation Experiments at NTCIR-6. In: Proceedings of NTCIR-6 (2007)
11. Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: SIGIR 1998, pp. 307–314 (1998)

Disambiguation and Unknown Term Translation in Cross Language Information Retrieval

Dong Zhou¹, Mark Truran², and Tim Brailsford¹

¹ School of Computer Science, University of Nottingham, United Kingdom

² School of Computing, University of Teesside, United Kingdom

dxz@cs.nott.ac.uk, m.a.truran@tees.ac.uk, tjb@cs.nott.ac.uk

Abstract. In this paper we present a report on our participation in the CLEF 2007 Chinese-English *ad hoc* bilingual track. We discuss a disambiguation strategy which employs a modified co-occurrence model to determine the most appropriate translation for a given query. This strategy is used alongside a pattern-based translation extraction method which addresses the ‘unknown term’ translation problem. Experimental results demonstrate that a combination of these two techniques substantially improves retrieval effectiveness when compared to various baseline systems that employ basic co-occurrence measures with no provision for out-of-vocabulary terms.

1 Introduction

Our participation in the CLEF 2007 *ad hoc* bilingual track was motivated by a desire to test and integrate two newly developed cross-language information retrieval (CLIR) techniques. The first of these techniques addresses the correct translation of *ambiguous* terms. A typical bilingual dictionary will provide a set of alternative translations for every term occurring within a given query. Choosing the correct translation for each term is a difficult procedure critical to the efficiency of any related retrieval functions. Previous solutions to this problem have employed co-occurrence information extracted from document collections to aid the process of resolving translation-based ambiguities [1], [2], [3], [4]. Extending this approach, we have developed a disambiguation strategy which employs a novel graph-based analysis of co-occurrence information to determine the most appropriate translations.

The second cross language IR technique we have developed addresses the *coverage problem*. Certain types of words are not commonly found in parallel texts or dictionaries, and it is these out-of-vocabulary (OOV) terms that will cause difficulties during automatic translation. Previous work on the problem of unknown terms has tended to concentrate upon complex statistical solutions [5]. We have adopted a much simpler approach to this problem which centres upon the application of linguistic and punctuative patterns to mixed language text [6].

Overall, the purpose of this paper is to evaluate a retrieval system which *combines* these two specific techniques in order to examine the effect of operating them concurrently.

2 Methodology

2.1 Resolution of Translation Ambiguities

The rationale behind the use of co-occurrence data to resolve translation ambiguities is relatively simple. For any query containing multiple terms which must be translated, the correct translations of individual query terms will tend to co-occur as part of a given sub-language, while the incorrect translations of individual query terms will not. Ideally, for each query term under consideration, we would like to choose the best translation that is consistent with the translations selected for all remaining query terms. However, this process of inter-term optimization has proved computationally complex for even the shortest of queries. A common workaround, used by several researchers working on this particular problem, involves use of an alternative resource-intensive algorithm, but this too has problems. In particular, it has been noted that the selection of translation terms is isolated and does not differentiate correct translations from incorrect ones [2], [4].

We approached this problem from a different direction. First of all, we viewed the co-occurrence of possible translation terms within a given corpus as *a graph*.

1. Input a query Q in a source language containing several terms $\{q_1, q_2, \dots, q_s\}$
2. For each term $q_i, i \in [1, s]$ in Q obtain the translation candidates $T(q_i) = \{t_{i,1}, t_{i,2}, \dots, t_{i,n_i}\}$
3. All possible translation candidates of the query terms are generated, to form a undirected weighted graph: $G = \langle F, W \rangle$, where F is the set of vertices representing one translation candidate $t_{i,j}$ to the query term q_i , and W is a complete set of *weighting functions*. Hence, every possible pairing of translation candidates sets has a corresponding weight attribute, w , which indicates the probability strength of a link potential between them. The set of weights is a whole can be denoted as:

$$W: F \times F \rightarrow \{w \in R: w \geq 0\}$$

An individual weighting between two translation candidates $t_{i,j}$ and $t_{k,l}$ is given by the function:

$$w(t_{i,j}, t_{k,l})$$

1. For each translation candidate, $t_{i,j}$, compute the *Centrality Score* and in order to determine: $CS(t_{i,j})$ for every single translation candidate in the graph.
3. The translation of a query term is then determined by selecting the translation candidate, $t_{i,j}$, which produces the max *Centrality Score* in the correspondent set of translation candidates:

$$t(q_i) = \underset{t_{i,j} \in T(q_i)}{Max}(t_{i,j})$$

5. Collect and output the final translation terms for the query Q .

Fig. 1. A graph-based algorithm for the disambiguation of query terms

In this graph, each translation candidate of a source query term is represented by a single node. Edges drawn between these nodes are weighted according to a particular co-occurrence measurement. We then apply graph-based analysis (inspired by research into hypermedia retrieval [7]) to determine the importance of a single node using global information recursively drawn from the entire graph. Subsequently, this measure of node importance is used to guide final query term translation.

The disambiguation algorithm we applied is summarized in Figure 1. The centrality score for a single vertex V_i in the graph is calculated in the following way: let $\{V_i\}_{in}$ be a set of nodes that point to V_i , and let $\{V_i\}_{out}$ be a set of nodes that V_i points at. Then, the centrality score of V_i is defined as follows:

$$Cen(V_i) = (1 - d)/N + d \times \sum_{j \in \{V_i\}_{in}} \frac{w_{i,j}}{\sum_{V_k \in \{V_j\}_{out}} w_{j,k}} Cen(V_j) \quad (1)$$

Where d is a dampening factor which integrates the probability of jumping from one node to another at random (normally set to 0.85) and N is the total number of nodes in the graph.

Two variations of the weighting function $w(t_{i,j} \leftrightarrow t_{k,l})$ have been developed. They are called *StrengthWeighting* (*SW*) and *FixedWeighting* (*FW*) respectively. The *SW* function should be considered an undirected weighted graph calculation while *FW* function is the undirected, unweighted alternative:

StrengthWeighting: If the similarity score (co-occurrence measurement) between two terms is more than zero, then the weight between the two terms is equal to the similarity score. Otherwise the weight is set to zero.

$$w(t_{i,j} \longleftrightarrow t_{k,l}) = \begin{cases} sim(t_{i,j}, t_{k,l}) & sim(t_{i,j}, t_{k,l}) > 0 \\ 0 & otherwise \end{cases}$$

FixedWeighting: If the similarity score (co-occurrence measurement) between two terms is more than zero, then the weight between the two terms is equal to one. Otherwise the weight is set to zero.

$$w(t_{i,j} \longleftrightarrow t_{k,l}) = \begin{cases} 1 & sim(t_{i,j}, t_{k,l}) > 0 \\ 0 & otherwise \end{cases}$$

2.2 Resolution of Unknown Terms

Our approach to the resolution of unknown terms is documented in detail in [6]. Stated succinctly, translations of unknown terms are obtained from a computationally inexpensive pattern-based processing of mixed language text retrieved from the web. A high level summary of this web based translation technique for OOV terms is as follows:

- The OOV term is submitted to a web search engine, and the results are cached.
- The text of each resultant web page is analyzed and certain punctuative and linguistic patterns are detected semi-automatically.

Category	Count	Examples	Obtained Translation
Name of Individuals	7	张瑞敏, 李连杰	zhangruimin, li lian jie
Name of Companies	1	联想	lianxiang co.
Name of Organizations	2	央视	cctv
Name of Places	2	五里乡	wuli
Verb	3	召开	zhao kai
Noun	0	战略	mo tai
Total	21		

Fig. 2. A breakdown of OOV terms found in the CLEF 2007 query set

- Analysis of detected patterns enables the identification of one or more translation candidates for the OOV term.
- A final translation for the OOV term is extracted from the list of candidates using a combination of extraction frequencies and pattern based weightings.

Illustrative examples of OOV terms translated using this approach can be found in Figure 2 (see also section 3):

3 Experiment

In the following section we describe our contribution to the CLEF 2007 *ad hoc* bilingual track. The document corpus used in our experiment was the English LA Times 2002 collection (135,153 English language documents) [8]. The queries we used were provided by the CLEF 2007 organizing committee and consisted of 50 multiple field topics written in Chinese, complete with relevance judgments.

3.1 Overview of the Experimental Process

A description of the CLIR process we adopted during this experiment is as follows: In the first step of the process we employed a naive bilingual dictionary to obtain a semi-translated query set¹. We then used the graph-based technique described above to resolve translation ambiguities within this set (with the co-occurrence scores obtained from the target document corpus). Finally, OOV terms occurring within the query set were passed to our pattern matcher to obtain final translation candidates. The fully translated queries were then passed to a information retrieval engine to retrieve the final document results list.

¹ <http://www ldc.upenn.edu/>

To prepare the corpus for the retrieval process, all of the documents were indexed using the Lemur toolkit². Prior to indexing, Porter’s stemmer [9] and a list of stop words³ were applied to the English documents.

3.2 Experimental Setup

In order to investigate the effectiveness of our various techniques, we performed a simple retrieval experiment with several key permutations. These variations are as follows:

MONO (monolingual): This part of the experiment involved retrieving documents from the test collection using Chinese queries manually translated into English by the CLEF 2007 organising committee. The performance of a monolingual retrieval system such as this has always been considered as an unreachable ‘upper-bound’ of CLIR as the process of automatic translation is inherently noisy.

ALLTRANS (all translations): Here we retrieved documents from the test collection using *all* of the translations provided by the bilingual dictionary for each query term.

FIRSTONE (first translations): This part of the experiment involved retrieving documents from the test collection using only the *first* translation suggested for each query term by the bilingual dictionary. Due to the way in which bilingual dictionaries are usually constructed, the first translation for any word generally equates to the most frequent translation for that term according to the World Wide Web.

COM (co-occurrence translation): In this part of the experiment, the translations for each query term were selected using the basic co-occurrence algorithm described in [3]. We used the target document collection to calculate the co-occurrence scorings.

GCONW (weighted graph analysis): Here we retrieved documents from the document collection using query translations suggested by our analysis of a weighted co-occurrence graph (i.e. we used the SW weighting function). Edges of the graph were weighted using co-occurrence scores derived using [3].

GCONUW (unweighted graph analysis): As above, we retrieved documents from the collection using query translations suggested by our analysis of the co-occurrence graph, only this time we used an unweighted graph (i.e. we used the FW weighting function).

GCONW+OOV (weighted graph analysis with unknown term translation): As GCONW, except that query terms that were not recognized (i.e. OOV terms) were sent to the unknown term translation system.

GCONUW+OOV (unweighted graph analysis with unknown term translation): As above, only this time we used the unweighted scheme.

3.3 Experimental Results and Discussion

The results of this experiment are provided in Tables 1 and 2. Document retrieval with no disambiguation of the candidate translations (ALLTRANS) was

² <http://www.lemurproject.org>

³ <ftp://ftp.cs.cornell.edu/pub/smart/>

Table 1. Short query results (*Title field only*)

	MAP	R-prec	P@10	% of MONO	IMPR. Over ALLTRANS	IMPR. over FIRSTONE	IMPR. over COM
MONO	0.4078	0.4019	0.486	N/A	N/A	N/A	N/A
ALLTRANS	0.2567	0.2558	0.304	62.95%	N/A	N/A	N/A
FIRSTONE	0.2638	0.2555	0.284	64.69%	2.77%	N/A	N/A
COM	0.2645	0.2617	0.306	64.86%	3.04%	0.27%	N/A
GCONW	0.2645	0.2617	0.306	64.86%	3.04%	0.27%	0.00%
GCONW+OOV	0.3337	0.3258	0.384	81.83%	30.00%	26.50%	26.16%
GCONUW	0.2711	0.2619	0.294	66.48%	5.61%	2.77%	2.50%
GCONUW+OOV	0.342	0.3296	0.368	83.86%	33.23%	29.64%	29.30%

Table 2. Long query results (*Title + Description fields*)

	MAP	R-prec	P@10	% of MONO	IMPR. Over ALLTRANS	IMPR. over FIRSTONE	IMPR. over COM
MONO	0.3753	0.3806	0.43	N/A	N/A	N/A	N/A
ALLTRANS	0.2671	0.2778	0.346	71.17%	N/A	N/A	N/A
FIRSTONE	0.2516	0.2595	0.286	67.04%	-5.80%	N/A	N/A
COM	0.2748	0.2784	0.322	73.22%	2.88%	9.22%	N/A
GCONW	0.2748	0.2784	0.322	73.22%	2.88%	9.22%	0.00%
GCONW+OOV	0.3456	0.3489	0.4	92.09%	29.39%	37.36%	25.76%
GCONUW	0.2606	0.2714	0.286	69.44%	-2.43%	3.58%	-5.17%
GCONUW+OOV	0.3279	0.3302	0.358	87.37%	22.76%	30.33%	19.32%

consistently the lowest performer in terms of mean average precision. This result was not surprising and merely confirms the need for an efficient process for resolving translation ambiguities.

When the translation for each query term was selected using a basic co-occurrence model (COM) [3], retrieval effectiveness always outperformed ALLTRANS and FIRSTONE. Interestingly, this result is inconsistent with earlier work published by [4] observing the opposite trend in the context of a TREC retrieval experiment.

Graph based analysis always outperformed the basic co-occurrence model (COM) on short query runs in both the weighted and un-weighted variants. However, COM scored higher than GCONW and GCONUW on runs with longer queries. This is probably due to the bilingual dictionary we selected for the experiment. The Chinese-English dictionary provided by LDC contains very few translation alternatives, thereby creating limited scope for ambiguity.

The combined model (graph based ambiguity resolution plus OOV term translation) scored highest in terms of mean average precision when compared to the non-monolingual systems. As illustrated by the data, improvement over the COM baseline is more pronounced for *Title* runs. This seems to reflect the length of

the queries. The *Title* fields in the CLEF query sets are very short, and correspondingly any OOV query terms not successfully translated will have a greater impact on retrieval effectiveness.

With respect to the monolingual system, a combination of our two new methods performed exceptionally well. For example, in the long query run, a combination of GCONW+OOV achieved 92.09% of monolingual performance. This means that our CLIR system as a whole achieved the second highest score in the whole CLEF 2007 *ad hoc* bilingual track (when compared with participating CLIR systems attempting retrieval of English documents using non-English query sets). This achievement is perhaps more remarkable when it is considered that our CLIR system does not yet employ *query expansion*, a technique renowned for improving retrieval effectiveness.

There were 21 unknown terms in the CLEF 2007 query set. Most of these terms were proper nouns or acronyms. Our system successfully translated 16 of the OOV terms, meaning its suggestions perfectly matched the manual CLEF 2007 translations. In our opinion, a translation hit rate of 76.2% in return for a meagre expenditure of resources emphatically validates the use of linguistics patterns in this context.

The OOV terms which were not successfully translated (23.8%) may have been *out of date*. Our method collects all translation candidates from the contemporaneous web. The query terms we worked with are several years old. It could be that the persons, organisations or acronyms which are referred to in that query set are no longer as prominent on the web as they once were. This would inevitably have a negative impact on our ability to generate appropriate translation candidates.

4 Conclusions

In this paper we have described our contribution to the CLEF 2007 Chinese-English *ad hoc* bilingual track. Our experiment involved the use of a modified co-occurrence model for the resolution of translation ambiguities, and a pattern-based method for the translation of OOV terms. The combination of these two techniques fared well, outperforming various baseline systems. The results that we have obtained thus far suggest that these techniques are far more effective when combined than in isolation.

Use of the CLEF 2007 document collection during this experiment has led to some interesting observations. There seems to be a distinct difference between this collection and the TREC alternatives commonly used by researchers in this field. Historically, the use of co-occurrence information to aid disambiguation has led to disappointing results on TREC retrieval runs [4]. Future work is currently being planned that will involve a side by side examination of the TREC and CLEF document sets in relation to the problems of translation ambiguity.

Acknowledgments. Thanks to people in WebTech Group in the University of Nottingham for many useful discussions. The work here was partially funded by a scholarship from the University of Nottingham and the Hunan University.

References

1. Ballesteros, L., Croft, W.B.: Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, pp. 64–71. ACM Press, New York (1998)
2. Gao, J., Nie, J.Y.: A study of statistical models for query translation: finding a good unit of translation. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA, pp. 194–201. ACM Press, New York (2006)
3. Jang, M.G., Myaeng, S.H., Park, S.Y.: Using mutual information to resolve query translation ambiguities and query term weighting. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland, pp. 223–229. Association for Computational Linguistics (1999)
4. Liu, Y., Jin, R., Chai, J.Y.: A maximum coherence model for dictionary-based cross-language information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, pp. 536–543. ACM Press, New York (2005)
5. Zhang, Y., Vines, P.: Using the web for automated translation extraction in cross-language information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom, pp. 162–169. ACM Press, New York (2004)
6. Zhou, D., Truran, M., Brailsford, T., Ashman, H.: Ntcir-6 experiments using pattern matched translation extraction. In: The sixth NTCIR workshop meeting, Tokyo, Japan, NII, pp. 145–151 (2007)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Ashman, H., Thistlewaite, P. (eds.) Proceedings of the 7th International World Wide Web Conference, vol. 30(1-7), pp. 107–117 (1998); reprinted In: Ashman, H., Thistlewaite, P.(eds.): Comput. Netw. ISDN Syst. 30(1-7), 107–117 (1998) 297827
8. Di Nunzio, G., Ferro, N., Mandl, T., Peters, C.: Clef 2007 ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)
9. Porter, M.F.: An algorithm for suffix stripping. Program 14, 130–137 (1980)

Cross-Language Retrieval with Wikipedia^{*}

Péter Schönhofen, András Benczúr, István Bíró, and Károly Csalogány

Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute, Hungarian Academy of Sciences
{schonhofen, benczur, ibiro, cskaresz}@ilab.sztaki.hu

Abstract. We demonstrate a twofold use of Wikipedia for cross-lingual information retrieval. As our main contribution, we exploit Wikipedia hyperlinkage for query term disambiguation. We also use bilingual Wikipedia articles for dictionary extension. Our method is based on translation disambiguation; we combine the Wikipedia based technique with a method based on bigram statistics of pairs formed by translations of different source language terms.

1 Introduction

In the paper we describe our cross-lingual retrieval (CLIR) method used in the Ad Hoc track of CLEF 2007 [1]. Novel in our approach is the exploitation of Wikipedia hyperlink structure for query term translation disambiguation; we also use bigram statistics for disambiguation baseline. Experiments performed on 250 Hungarian and 550 German source language topics against the English target collections GH, LAT94 and LAT02 show 1–3% improvement in MAP by using Wikipedia. The MAP of translated queries was roughly 62% of the original ones for Hungarian and 80–88% for German source language queries when measured over the $TF \times IDF$ -based Hungarian Academy of Sciences search engine [2].

Due to the morphological complexity, ad Hoc retrieval in Hungarian is a hard task with performances in general reported below those of the major European languages (French or Portuguese from CLEF 2006) [3,4,5]. Good quality overview resources on the Hungarian grammar can be found on the Wikipedia [6].

Our CLIR method falls in the branch of machine translation based on disambiguation between multiple possible dictionary entries [7]. In our method we first generate a raw word-by-word translation of the topic narrative by using the online dictionary of the Hungarian Academy of Sciences¹. In the disambiguation phase we keep at least one translation for one word but discard off-topic translations. First we disambiguate translations of pairs of words by the bigram language model of Wikipedia, then we score remaining candidates by mapping them to Wikipedia articles and analyzing Wikipedia hyperlinks between them.

When using Wikipedia linkage for query term disambiguation, our fundamental idea is to score candidate English terms based on the strength of their semantical relation to

^{*} This work was supported by a Yahoo! Faculty Research Grant and by grants MOLINGV NKFP-2/0024/2005, NKFP-2004 project Language Miner <http://nyelvbanyasz.sztaki.hu>

¹ <http://dict.sztaki.hu/>

other candidates. Our method is a simplified version of translation disambiguation that typically also involves the grammatical role of the source phrase (e.g. [8]), an information unavailable for a typical query phrase.

Several researcher utilized ontologies to support disambiguation in machine translation [9], or as a base for internal representation bridging the source and target languages [10]; [11] provides an extensive theoretical discussion on this topic. However, due to the lack of ontologies which have a sufficiently wide coverage and at the same time are available in multiple languages, these methods typically construct their own ontologies through some machine learning technique over parallel corpora. Though the idea of taking advantage of Wikipedia has already emerged, either as an ontology [12] or as a parallel corpus [13], to our best knowledge, so far it has not been used for dictionary construction or to improve translation accuracy.

2 Our Method

Our CLIR method consists of a word-by-word translation by dictionary, then a two-phase term translation disambiguation, first by bigram statistics, then, as our novel idea, by exploiting the Wikipedia hyperlink structure. We use the same stemming and stop-word removal procedure for the target corpus, the dictionaries and Wikipedia articles; note that this may result in merging dictionary or Wikipedia entries. Stemming for English and German is performed by TreeTagger [14] and for Hungarian by an open source stemmer [4].

For dictionary we use an online dictionary as well as bilingual Wikipedia articles. The Hungarian Academy of Sciences online dictionary consists of an official and a community edited sub-dictionary, comprising of roughly 131,000 and 53,000 entries, respectively. We extend the dictionary by linked pairs of English and Hungarian Wikipedia article titles, a method that currently only slightly increases coverage since as of late 2007, there are only 60,000 articles in Hungarian Wikipedia mostly covering technical terms, geographical locations, historical events and people either not requiring translation or rarely occurring inside query text. We order translations by reliability and discard less reliable translations even if they would provide additional English translations for a source language term. The reliability order is official dictionary, community edited dictionary and finally translations generated from bilingual Wikipedia.

For German as source language we used the SMART English-German plugin version 1.4 [15] and the German and English Wiktionaries (which for many phrases specify the corresponding English and German terms, respectively, in their “Translations” section) as dictionaries. In addition, we collected translations from the titles of English and German Wikipedia articles written about exactly the same topic (and explicitly linked to each other through the cross-language facility of Wikipedia). We worked with the snapshots of Wiktionary and Wikipedia taken in September, 2007. Reliability ranking of the dictionaries is as follows: the SMART plugin (89,935 term pairs), followed by Wiktionary (9,610 new term pairs), and bilingual Wikipedia (126,091 new term pairs).

As we can see from the example of dictionary translations for some Hungarian words shown in Table 1, the dictionary typically gives a relatively large number of possible translations, whose majority evidently belongs to the wrong concept.

Table 1. Possible translations of Hungarian words from queries #251 (alternative medicine), #252 (European pension schemes) and #447 (politics of Pym Fortuin) along with their bigram and Wikipedia disambiguation scores. Sorting is by the combined score (not shown here).

Hungarian word	Bigram score	Wikipedia score	English word
természetes	0.0889	0.1667	natural
	0.3556	0.0167	grant
	0.0000	0.1833	natural ventilation
	0.0000	0.0167	naturalism
	0.0000	0.0167	genial
	0.0000	0.0167	naturalness
	0.0000	0.0167	artless
	0.0000	0.0010	naivete
	0.0000	0.0010	unaffected
kor ² (kőr)	0.2028	0.2083	age
	0.1084	0.1250	estate
	0.0385	0.0833	period
	0.0105	0.0625	cycle
	0.0350	0.0208	era
	0.0000	0.0625	epoch
	0.0000	0.0052	asl
	0.0000	0.0010	temp
ellentmondás	0.0090	0.1961	paradox
	0.0601	0.0589	conflict
	0.0060	0.0392	contradiction
	0.0030	0.0196	variance
	0.0060	0.0098	discrepancy
	0.0060	0.0049	contradict
	0.0060	0.0049	inconsistency

First we disambiguate by forming all possible pairs E, E' of English translations of different source language terms present in the same paragraph (query title, description or narratives). We precompute bigram statistics of the target corpus. We let $\text{rank}_B(E)$ be the maximum of all bigram counts with other terms E' . See the second column of Table 1 for typical scores.

Our main new idea is the second disambiguation step that uses Wikipedia transformed into a concept network based on the assumption that reference between articles indicates semantic relation. As opposed to proper ontologies such as WordNet or OpenCyc, here relations have no type (however, several researchers worked out techniques to rectify this omission, for instance [16]). Note however that Wikipedia itself is insufficient for CLIR itself since it deals primarily with complex concepts while basic nouns and verbs (e.g. “read”, “day”) are missing and hence we use it in combination with bigrams.

We preprocess Wikipedia in a way described in [17]; there the way special pages such as redirects, category pages etc. are handled is described in detail. We used the

² Term “kor” has different meanings (age, illness, cycle, heart suit) depending on the diacritics (see in brackets).

Wikipedia snapshot taken in August of 2006 with (after preprocessing) 2,184,226 different titles corresponding to 1,728,641 documents.

We label query terms by Wikipedia documents by Algorithm 1. First we find Wikipedia title words in the query; for multiword titles we require an exact matching sequence of the translated topic definition. In this way we obtain a set W_E of Wikipedia titles corresponding to translated words E . The final labels arise as the top ranked concepts after a ranking procedure that measures connectivity within the graph of the concept network as in Algorithm 1. In the algorithm first we rank Wikipedia documents D by the number of links to terms O in the source language, i.e. the number of such O with a translation E' that has a $D' \in W_{E'}$ linked to D . For each translation E we then take the maximum of the ranks within W_E . We add these ranks up in the case of multiword translations.

Algorithm 1. Outline of the labeling algorithm

```

for all English translation words  $E$  do
  for all Wikipedia documents  $D$  with title  $T_D$  containing  $E$  do
    if  $T_D$  appears as a sequence around  $E$  then
      add  $D$  to the list  $W_E$ 
  for all translation words  $E'$  do
    for all Wikipedia documents  $D \in W_{E'}$  do
       $\text{rank}_W(D) \leftarrow |\{O : O \text{ is a source language word such that there is a translation } E' \text{ and}$ 
         $\text{a } D' \in W_{E'} \text{ with a link between } D \text{ and } D' \text{ in Wikipedia}\}|$ 
     $\text{rank}(E) \leftarrow \max\{\text{rank}(D) : D \in W_E\}$ 
  
```

The third column of Table 1 shows scores of various English translation candidates computed from the degree of linkage between their corresponding Wikipedia article(s) and those of other candidates.

Finally we build the query based on the bigram and Wikipedia based ranks rank_B and rank_E of the individual translations, by also taking the quality $q(E)$ of the dictionary that contains the translation into account. We choose the translation that maximizes the expression below; in cases of ties we keep both:

$$q(E) \cdot (\log(\text{rank}_B(E)) + \alpha \log(\text{rank}_W(E))). \quad (1)$$

In Table 1, translations are ordered according to their combined scores. Note that for the first word, neither bigram statistics, nor Wikipedia would rank the correct translation to the first place, but their combined score does. For the second word, both scoring would select the same candidate as the best one, and for the third, Wikipedia scoring is right while the bigram statistics is wrong.

3 Search Engine

We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system that uses a $TF \times IDF$ -based ranking in a weighted combination of the following factors:

Table 2. Topics used in the CLIR experiments

source language	English	German	Hungarian
GH	141–350	141–350	251–350
LAT 94	1–350	1–200, 250–350	251–350
LAT 02	401–450		401–450

- Proximity of query terms as in [18,19];
- Document length normalization [20];
- Different weights to different parts of the document (title, or location as in case of ImageCLEF-Photo topics);
- Total weight of query terms in the document; the original query is considered as a weighted OR query with reduced weights to words in description and narrative.
- The proportion of the document between the first and last query term, a value almost 1 if the document contains query terms at the beginning and at the end, and $1 / \text{size}$ for a single occurrence.

We observed that we get the best result if the weight of the number of query terms is much higher than the $TF \times IDF$ score. In other words, we rank documents with respect to the number of query terms found inside their text, then use the $TF \times IDF$ -based measurement to differentiate between documents carrying the same number of query terms.

We translate all of title, description and narrative that we all use for recognizing the concept of the query mapped into Wikipedia. For retrieval we then use translations with weights 1 for title, 0.33 for description and 0.25 for narrative.

4 Results

Retrieval performance for Hungarian as source language for the CLEF 2007 topics 401–450 is shown in Table 3. In addition we use a wide range of topics listed in Table 2 to show average performance for both Hungarian and German as source language in Table 4. We use the English topics as baseline; we also show performance for the bigram based and the combined bigram and Wikipedia translation disambiguation steps.

Table 5 shows (the first four words of) sample queries where the Wikipedia-enhanced translation is much less or much more effective than the official English translation. Mistakes corrected by Wikipedia-based disambiguation can be classified in five main groups:

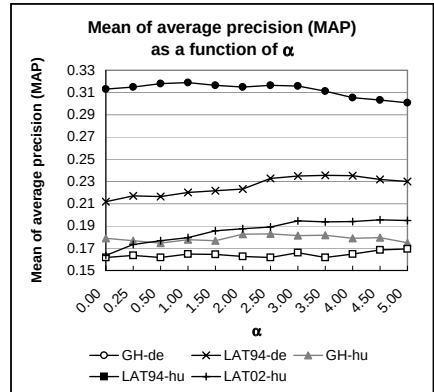
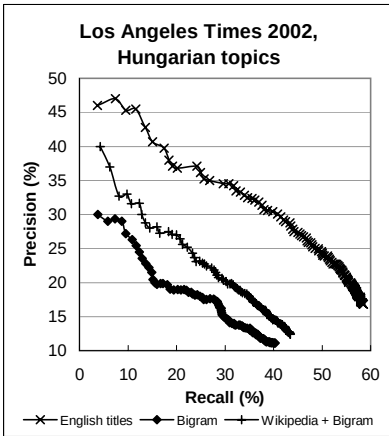
- translated words are the same but are in a different grammatical form (see for example “human cloning” vs. “human clone” in topic 408);
- translated words are synonyms of the original ones (e.g. “Australian” vs. the informal “Aussie” in topic 407);
- insufficient information to properly disambiguate (topic 409);
- the Hungarian stemmer failed to determine the stem of a word (leaving the Hungarian word untranslated for “drug” in topic 415);
- the dictionary failed to provide any translation for a given word (see for instance “weapon” in topic 410, which should have been recognized indirectly from a Hungarian compound term).

Table 3. Performance over Hungarian topics of Table 2, including CLEF 2007 topics (LAT 02)

Corpus	method	P @ 5	R @ 5	P @ 10	R @ 10	MRR	MAP
GH	English topics	34.46	21.70	28.43	30.62	0.5304	0.2935
	Bigram	21.20	10.40	20.24	18.74	0.3857	0.1666
	Bigram + Wikipedia	23.86	12.94	21.45	20.98	0.4038	0.1826
LAT 94	English topics	33.90	16.81	28.74	24.00	0.5245	0.2638
	Bigram	20.21	11.55	17.26	16.74	0.3632	0.1515
	Bigram + Wikipedia	20.00	12.98	18.32	18.74	0.3830	0.1693
LAT 02	English topics	42.80	13.61	36.80	20.20	0.6167	0.2951
	Bigram	27.20	9.55	22.80	13.68	0.4518	0.1566
	Bigram + Wikipedia	31.60	10.68	28.20	15.95	0.5348	0.1956

Table 4. Performance over German topics of Table 2

Corpus	method	P @ 5	R @ 5	P @ 10	R @ 10	MRR	MAP
GH	English topics	34.12	27.78	27.00	36.66	0.5594	0.3537
	Bigram	28.00	24.21	22.71	32.88	0.4940	0.3011
	Bigram + Wikipedia	29.41	25.99	22.88	33.62	0.5078	0.3188
LAT-94	English topics	36.10	19.65	30.04	27.97	0.5746	0.2974
	Bigram	27.56	15.36	22.80	22.29	0.4667	0.2055
	Bigram + Wikipedia	29.43	18.34	24.39	24.98	0.4863	0.2327

**Fig. 1.** Left: Precision–recall curve for the CLEF 2007 Hungarian topics when using the original English titles as well as bigram based and combined disambiguation. Right: Effect of α on the mean of average precision (MAP)

Wikipedia based translations are usually more verbose than raw translations by introducing synonyms (like in topic 412) but also sometimes strange words (such as in topic 414). We often reintroduce important keywords lost in bigram based disambiguation as e.g. “cancer” in topic 438, “priest” in topic 433. As a result, though precision at

5 retrieved documents were only 27.20% for raw translations, a fraction of the 42.80% observed when using official English translations, Wikipedia post-processing managed to increase precision to 31.60%. Figure 1 shows the precision–recall curve as well as how the bigram and Wikipedia based disambiguation combination weight factor α as in (1) affects average precision.

Table 5. Sample queries where Wikipedia based disambiguation resulted in improved (above) and deteriorated (below) average precision over bigram based disambiguation, sorted by difference

Topic No.	Avg. prec. (En. titles)	Avg. prec. (Wikiped.)	Difference	English title	Wikipedia-enhanced translation
404	0.0667	1.0000	-0.9333	nato summit security	safety security summit nato ...
408	0.0890	0.4005	-0.3115	human cloning	human clone number statement ...
412	0.0640	0.2077	-0.1437	book politician	book politician collection anecdote ...
409	0.3654	0.4997	-0.1343	bali car bombing	car bomb bali indonesia ...
421	0.3667	0.4824	-0.1157	kostelic olympic medal	olympic kostelic pendant coin ...
438	0.0071	0.0815	-0.0744	cancer research	oncology prevention cancer treatment ...
425	0.0333	0.1006	-0.0673	endangered species	endanger species illegal slaughter ...
406	0.0720	0.1314	-0.0594	animate cartoon	cartoon award animation score ...
432	0.3130	0.3625	-0.0495	zimbabwe presidential election	presidential zimbabwe marcius victor ...
417	0.0037	0.0428	-0.0391	airplane hijacking	aircraft hijack diversion airline ...

Topic No.	Avg. prec. (En. titles)	Avg. prec. (Wikiped.)	Difference	English title	Wikipedia-enhanced translation
407	0.7271	0.0109	0.7162	australian prime minister	premier aussie prime minister ...
448	0.7929	0.1691	0.6238	nobel prize chemistry	nobel chemistry award academic ...
416	0.6038	0.0000	0.6038	moscow theatre hostage crisis	moscow hostage crisis theatre ...
410	0.6437	0.0947	0.5490	north korea nuclear weapon ...	north korean korea obligation ...
402	0.4775	0.0056	0.4719	renewable energy source	energy parent reform current ...
414	0.4527	0.0000	0.4527	beer festival	hop festival good line ...
441	0.6323	0.1974	0.4349	space tourist	tourist space russian candidate ...
443	0.5022	0.1185	0.3837	world swimming record	swim time high sport ...
427	0.6335	0.2510	0.3825	testimony milosevic	milosevic testimony versus hague ...
419	0.3979	0.0233	0.3746	nuclear waste repository	waste atom cemetery federal ...
401	0.4251	0.0899	0.3352	euro inflation	price rise euro introduction ...

References

1. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)
2. Benczúr, A.A., Csalogány, K., Fogaras, D., Friedman, E., Sarlós, T., Uher, M., Windhager, E.: Searching a small national domain – A preliminary report. In: Proceedings of the 12th International World Wide Web Conference (WWW) (2003)
3. Di Nunzio, G., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
4. Halácsy, P., Trón, V.: Benefits of deep NLP-based lemmatization for information retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
5. Savoy, J., Abdou, S.: UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)

6. Hungarian Grammar: From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Hungarian_grammar
7. Hiemstra, D., de Jong, F.: Disambiguation strategies for cross-language information retrieval. In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, London, UK, pp. 274–293 (1999)
8. Dorr, B.J.: The use of lexical semantics in interlingual machine translation. *Machine Translation* 7(3), 135–193 (1992)
9. Knight, K., Luk, S.K.: Building a large-scale knowledge base for machine translation. In: Proceedings of the twelfth National Conference on Artificial Intelligence, pp. 773–778 (1994)
10. Navigli, R., Velardi, P., Gangemi, A.: Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems* 18(1), 22–31 (2003)
11. Mahesh, K.: Ontology development for machine translation: Ideology and methodology. Technical Report MCCS 96-292, Computing Research Laboratory, New Mexico State University (1996)
12. Denoyer, L., Gallinari, P.: The Wikipedia XML corpus. *SIGIR Forum* 40(1), 64–69 (2006)
13. Adafre, S.F., de Rijke, M.: Finding similar sentences across multiple languages in Wikipedia. In: Proceedings of the New Text Workshop, 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)
14. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing (1994)
15. Project jDictionary: SMART English-German plugin version 1.4, <http://jdictionary.sourceforge.net/plugins.html>
16. Völkel, M., Kröttsch, M., Vrandečić, D., Haller, H., Studer, R.: Semantic Wikipedia. In: Proceedings of the 15th international conference on World Wide Web, pp. 585–594 (2006)
17. Schönhofen, P.: Identifying document topics using the Wikipedia category network. In: *Web Intelligence*, pp. 456–462 (2006)
18. 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)
19. Büttcher, S., Clarke, C.L.A., Lushman, B.: Term proximity scoring for Ad-Hoc retrieval on very large text collections. In: *SIGIR 2006*, pp. 621–622. ACM Press, New York (2006)
20. Singhal, A., Buckley, C., Mitra, M., Salton, G.: Pivoted document length normalization. Technical Report TR95-1560, Cornell University, Ithaca, NY (1995)

Cross-Lingual Information Retrieval System for Indian Languages

Jagadeesh Jagarlamudi and A. Kumaran

Multilingual Systems Research,
Microsoft Research India,
Bangalore, India

{jags,a.kumaran}@microsoft.com
<http://research.microsoft.com/india>

Abstract. This paper describes our attempt to build a Cross-Lingual Information Retrieval (CLIR) system as a part of the Indian language sub-task of the main Adhoc monolingual and bilingual track in CLEF¹ competition. In this track, the task required retrieval of relevant documents from an English corpus in response to a query expressed in different Indian languages including Hindi, Tamil, Telugu, Bengali and Marathi. Groups participating in this track were required to submit a English to English monolingual run and a Hindi to English bilingual run with optional runs in rest of the languages. Our submission consisted of a monolingual English run and a Hindi to English cross-lingual run.

We used a word alignment table that was learnt by a Statistical Machine Translation (SMT) system trained on aligned parallel sentences, to map a query in the source language into an equivalent query in the language of the document collection. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. On the CLEF 2007 data set, our official cross-lingual performance was 54.4% of the monolingual performance and in the post submission experiments we found that it can be significantly improved up to 76.3%.

1 Introduction

The rapidly changing demographics of the internet population [1] and the plethora of multilingual content on the web [2] has attracted the attention of the Information Retrieval(IR) community to develop methodologies for information access across languages. Since the past decade [3,4,5,6], researchers are looking at ways to retrieve documents in a language in response to a query in another language. This fundamentally assumes that users can read and understand documents written in foreign language but are unable to express their information need in that language. There are arguments against this assumption as well: for example, Moulinier and Schilder argue that it is unlikely that the information in another language will be useful unless users are already fluent in that language [7]. However, we argue that in specific cases such methodologies could still be

¹ Cross Language Evaluation Forum, <http://www.clef-campaign.org>

valid. For example, in India students learn more than one language from their childhood and more than 30% of the population can read and understand Hindi apart from their native language [8]. The multilingual capability of the users exhibits a great demand for systems with the capability to retrieve relevant documents in languages different from the language in which information need is expressed.

Lack of resources is still a major reason for relatively less number of efforts in the cross-lingual setting in the Indian subcontinent. Research communities working in Indian Languages, especially on Machine Translation (MT) [9], have built some necessary resources like morphological analyzer and bilingual dictionaries for some languages. As we demonstrate in this paper, even though these resources are built mainly for MT, they can still be used as a good starting point to build a CLIR system. More specifically, in this paper we will describe our first attempt at building a CLIR system using a bilingual statistical dictionary that was learnt automatically during the training phase of a SMT [10] system.

In the rest of the paper we will first define the problem in section 2, followed by a brief description of our approach in section 3. In section 4 we will describe data set along with the resources used and also present the results reported in the CLEF competition (sec. 4.1). This section also includes some analysis of the results. Section 5 presents conclusion and outlines our plans for future work.

2 Problem Statement

We have participated in the Indian Language sub-task of the main CLEF 2007 Ad-hoc monolingual and bilingual track. This track tests the performance of systems in retrieving relevant documents in response to a query in the same and different language from that of the document set. In the Indian language track, documents are provided in English and queries are specified in different languages including Hindi, Telugu, Bengali, Marathi and Tamil. The system has to retrieve 1000 relevant documents as a response to a query in any of the above mentioned languages. All the systems participating in this track are required to submit an English to English monolingual run and a Hindi to English bilingual run. Runs in the rest of the languages are optional. We have submitted an English to English monolingual and Hindi to English bilingual run.

3 Approach

Converting the information expressed in different languages to a common representation is inherent in cross-lingual applications to bridge the language barrier. In CLIR, either the query or the document or both need to be mapped onto the common representation to retrieve relevant documents. Translating all documents into the query language is less desirable due to the enormous resource requirements. Usually the query is translated into the language of the target collection of documents. Typically three types of resources are being exploited for translating the queries: bilingual machine readable dictionaries, parallel texts

and machine translation systems. MT systems typically produce one candidate translation thus some potential information which could be of use to IR system is lost. Even though researchers [11] have also explored considering more than one possible translation to avoid the loss of such useful information, another difficulty in using the MT system comes from the fact that most of the search queries are very short and thus lack necessary syntactic information required for translation. Hence most approaches use bilingual dictionaries.

In our work, we have used statistically aligned Hindi to English word alignments that were learnt during the training phase of a machine translation system. The query in Hindi is translated into English using word by word translation. For a given Hindi word, all English words which have translation probability above a certain threshold are selected as candidate translations. Only top ' n ' of these candidates are selected as final translations in order to reduce ambiguity in the translation. This process may not produce any translations for some of the query words because either the word is not available in the parallel corpus or all translation probabilities are less than the threshold. In such cases, we attempt to transliterate the query word into English. We have used a noisy channel model based transliteration algorithm [12]. The phonemic alignments between Hindi characters and corresponding English characters are learnt automatically from a training corpus of parallel names in Hindi and English. These alignments along with their probabilities are used, during viterbi decoding, to transliterate a new Hindi word into English. As reported, this system will output the correct (fuzzy match) English word in top 10 results, with an accuracy of about 30%(80%). As a post processing step, target language vocabulary along with approximate string matching algorithms like soundex and edit distance measure [13] were used to filter out the correct word from the incorrect ones among the possible candidate transliterations.

Once the query is translated into the language of the document collection, standard IR algorithms can be used to retrieve relevant documents. We have used Language Modeling [14] in our experiments. In a Language Modeling framework, both query formulation and retrieval of relevant documents are treated as simple probability mechanisms. Essentially, each document is assumed to be a language sample and the query to be a sample from the document. The likelihood of generating a query from a document ($p(q|d)$) is associated with the relevance of the document to the query. A document which is more likely to generate the user query is considered to be more relevant. Since a document, considered as a bag of words is very small, compared to the whole vocabulary, most of the times the resulting document models are very sparse. Hence smoothing of the document distributions is very crucial. Many techniques have been explored and because of its simplicity and effectiveness we chose the relative frequency of a term in the entire collection to smooth the document distributions.

In a nutshell, structural query translation [6] is used to translate query into English. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. The following section describes our approach applied in the CLEF 2007 participation and some further experiments to calibrate the quality of our system.

4 Experiments

In both the Adhoc bilingual ‘X’ to English track and Indian language sub track, the target document collection consisted of 135,153 English news articles published in Los Angeles Times, from the year 2002. During indexing of this document collection, only the text portion (embedded in <LD> and <TE> tags) was considered. Note that the results reported in this paper do not make use of other potentially useful information present in the document, such as, the document heading (with in <DH> tag) and the photo caption (in <CP> tag), even though we believe that including such information would improve the performance of the system. The resulting 85,994 non-empty documents were then processed to remove stop words and the remaining words were reduced into their base form using Porter stemmer [15].

The query set consists of 50 topics originally created in English and translated later into other languages. For processing Hindi queries, a list of stop words was formed based on the frequency of a word in the monolingual corpus corresponding to the Hindi part of the parallel data. This list was then used to remove any less informative words occurring in the topic statements. The processed query was then translated into English using a word alignment table.

We have used a word alignment table that was learnt by the SMT [10] system trained on 100K Hindi to English parallel sentences acquired from Webdunia to translate Hindi queries. Since these alignments were learnt for machine translation purpose, the alignments included words in their inflectional forms. For this reason we have not converted the query words into their base form during the translation. Table 1 shows the statistics about the coverage of the alignment table corresponding to different levels of threshold on the translation probability (column 1), note that a threshold value of 0 corresponds to having no threshold at all. Columns 2 and 3 indicate the coverage of the dictionary in terms of source and target language words. The last column denotes the average number of English translations for a Hindi word. It is very clear and intuitive that as the threshold increases the coverage of the dictionary decreases. It is also worth noting that as the threshold increases the average translations per source word decreases, indicating that the target language words which are related to the source word but not synonymous are getting filtered.

4.1 Results

For each query, a pool of candidate relevant documents is created by combining the documents submitted by all systems. From this pool assessors filter out actual

Table 1. Coverage statistics of the word alignment table

Threshold	Hindi words	English words	Translations per word
0	57555	59696	8.53
0.1	45154	54945	4.39
0.3	14161	17216	1.59

Table 2. Monolingual and Cross-lingual experiments

	Monolingual		Crosslingual	
	LM(td)	LM(tdn)	LM(td)	LM(tdn)
MAP	0.3916	0.3964	0.1994	0.2156
p@10	0.456	0.454	0.216	0.294

relevant documents from the non relevant ones. These relevance judgements are then used to automatically evaluate the quality of participating CLIR systems.

Here we present the official results of our monolingual English run and Hindi to English bilingual run. In our case, we specifically participated in only one Indian language - Hindi, though the data was available in 5 Indian languages. For our official submission, with the aim of reducing noise in the translated query, we used a relatively high threshold of 0.3 for the translation probability. To avoid ambiguity, when there are many possible English translations for a given Hindi word, we included only the two best possible translations according to the translation table. Table 2 shows the official results of our submission. We have submitted different runs using title, description (td) and title, description and narration (tdn) as query.

In a second set of experiments, we experimented with the effect of various levels of threshold and the number of translations above the threshold on Mean Average Precision(MAP) score. The results obtained by the cross-lingual system with varying threshold are compared against the monolingual system in fig. 1. The right most bar in each group represents the monolingual performance of the system. The figure shows that in each group the performance increases as the threshold decrease and it decreases if you consider more number of

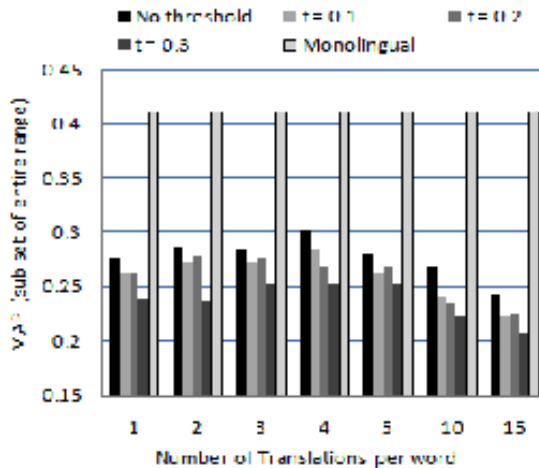


Fig. 1. Hindi to English cross-lingual performance with varying levels of threshold on translation probability

Table 3. Fraction of translated words

Threshold	Translated words
0	0.803
0.1	0.7892
0.2	0.711
0.3	0.6344

Table 4. Skewedness of the dictionary

No. of top words (% of vocab size)	% of dictionary entries
10 (0.0173)	3.44
20 (0.0347)	5.87
50 (0.0869)	11.05
100 (0.173)	17.01

possible translations per word (drop when 10 and 15 translations were considered) perhaps due to the shift in query focus with the inclusion of many less synonymous target language words. For the CLEF data set, we found that considering the four most possible translations without any threshold (left most bar) of the translation probability gave us the best results (73.4% of monolingual IR performance).

As the threshold decreases potentially two things can happen: words which were not translated previously can get translated or new target language words whose translation probability was below the threshold earlier will now become part of the translated query. Table 3 shows the fraction of query words that were translated corresponding to each of these thresholds. Table 3 and figure 1 show that as the threshold on the translation probability decreases, the fraction of query words getting translated increases, resulting in an overall increase in system performance. But the performance increase between having no threshold and a threshold of 0.1 compared to the small fraction of new words that got translated suggest that even noisy translations, even though they are not truly synonymous, might help CLIR. This is perhaps due to the fact that for the purposes of IR, in absense of appropriate translation having a list of associated words may be sufficient to identify the context of the query [16].

During the query wise analysis we found that cross-lingual retrieval performed marginally better than the monolingual system on five topics, whereas it under performed on 19 queries. The analysis of 11 queries, on which the CLIR performance is worse than 10% of monolingual, revealed three kinds of errors as very prominent: words missing in the translation dictionary, inappropriate selection of target language word and transliteration errors. Even though the dictionary coverage looks very exhaustive, queries still have relatively more number of content words which are not covered by the dictionary. This is mainly because the dictionary since it is learned statistically, it is skewed. Table 4 shows the number of top words, in terms of having multiple target language translations, along with the fraction of dictionary entries corresponding to them. The top 100 words

corresponding to a small fraction of 0.173% of vocabulary cover almost 17% of the dictionary and most of these words are less informative for Information Retrieval purpose. This is evident as the dictionary is learnt statistically, where a frequent word has more chances of being aligned to many target language words depending on the context in which it is being used.

If a query contains a relatively more frequent word, since it has more chances of being aligned to many words, it may become a source of noise in the translated query. This has also been the cause for the selection of inappropriate target language words, even when you consider only top 'n' translations. In order to select the appropriate target language word that is more common in the target language corpus, we performed a simple adaptation of the translation dictionary. Instead of using $p(e|h)$ directly we have replaced it with $p(h|e) \cdot p(e)$ and normalized appropriately to follow the probability constraints. The unigram probability counts for the English words are computed from the target language corpus instead of the English part of parallel sentences. We assumed that this will prefer words that are more frequent in the collection to words that are relatively infrequent. Such a simple technique has resulted in an improvement of 4.23% indicating a scope for further exploration.

5 Conclusion and Future Work

This paper describes our first attempt at building a CLIR system with the help of a word alignment table learned statistically. We present our submission in the Indian language sub-task of the Adhoc monolingual and bilingual track of CLEF 2007. In post submission experiments we found that, on CLEF data set, a Hindi to English cross-lingual information retrieval system using a simple word by word translation of the query with the help of a word alignment table was able to achieve $\sim 76\%$ of the performance of the monolingual system. Empirically we found that considering four most probable word translations with no threshold on the translation probability gave the best results.

In our analysis, we found the coverage of dictionary and the choice of appropriate translation to be the potential places for improvement. In the future we would like to exploit either the parallel or comparable corpora for selecting the appropriate translation of a given source word. We would also like to compare the distribution statistics of a statistically learned dictionary with respect to a hand crafted dictionary of similar size to compare them for CLIR purposes.

References

1. Internet, <http://www.internetworldstats.com>
2. GlobalReach, <http://www.global-reach.biz/globstats/evol.html>
3. Ballesteros, L., Croft, W.B.: Dictionary methods for cross-lingual information retrieval. In: Thoma, H., Wagner, R.R. (eds.) DEXA 1996. LNCS, vol. 1134, pp. 791–801. Springer, Heidelberg (1996)

4. Hull, D.A., Grefenstette, G.: Querying across languages: A dictionary-based approach to Multilingual Information Retrieval. In: SIGIR 1996: Proc. of the 19th annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 49–57. ACM Press, New York (1996)
5. McNamee, P., Mayfield, J.: Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In: SIGIR 2002: Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 159–166. ACM Press, New York (2002)
6. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval* 4(3-4), 209–230 (2001)
7. Moulinier, I., Schilder, F.: What is the future of multi-lingual information access?. In: SIGIR 2006 Workshop on Multilingual Information Access 2006, Seattle, Washington, USA (2006)
8. Burkhart, G.E., Goodman, S.E., Mehta, A., Press, L.: The Internet in India: Better times ahead?. *Commun. ACM* 41(11), 21–26 (1998)
9. Bharati, A., Sangal, R., Sharma, D.M., Kulakarni, A.P.: Machine Translation activities in India: A survey. In: Workshop on survey on Research and Development of Machine Translation in Asian Countries (2002)
10. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. *Comput. Linguist.* 29(1), 19–51 (2003)
11. Kwok, K.L., Choi, S., Dinstl, N.: Rich results from poor resources: Ntcir-4 monolingual and cross-lingual retrieval of korean texts using chinese and english. *ACM Transactions on Asian Language Information Processing (TALIP)* 4(2), 136–162 (2005)
12. Kumaran, A., Kellner, T.: A generic framework for machine transliteration. In: SIGIR 2007: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 721–722. ACM Press, New York (2007)
13. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: *English Translation in Soviet Physics Doklady*, pp. 707–710 (1966)
14. Ponte, J.M., Croft, W.B.: A Language Modeling Approach to Information Retrieval. In: *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–281 (1998)
15. Porter, M.F.: An algorithm for suffix stripping. *Program: News of Computers in British University libraries* 14, 130–137 (1980)
16. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. *Inf. Process. Manage.* 43(4), 866–886 (2007)

Bengali, Hindi and Telugu to English Ad-Hoc Bilingual Task at CLEF 2007

Sivaji Bandyopadhyay, Tapabrata Mondal, Sudip Kumar Naskar,
Asif Ekbal, Rejwanul Haque, and Srinivasa Rao Godhavarthy

Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032, India
sivaji_cse_ju@yahoo.com, sbandyopadhyay@cse.jdvu.ac.in

Abstract. This paper presents the experiments carried out at Jadavpur University as part of the participation in the CLEF 2007 ad-hoc bilingual task. This is our first participation in the CLEF evaluation task and we have considered Bengali, Hindi and Telugu as query languages for the retrieval from English document collection. We have discussed our Bengali, Hindi and Telugu to English CLIR system as part of the ad-hoc bilingual task, the English IR system for the ad-hoc monolingual task and the associated experiments at CLEF. Query construction was manual for Telugu-English ad-hoc bilingual task, while it was automatic for all other tasks.

1 Introduction

Cross-language information retrieval (CLIR) research involves the study of 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. Cross-language information retrieval focus on the cross-language issues from information retrieval (IR) perspective rather than machine translation perspective.

Many different techniques were tested in various CLIR systems in the past in order to address the issues of transferring the query term from the source to the target language, the mechanisms of determining the possible translations from the source to the target language and the methods of weighing the different translations alternatives. These techniques can be broadly classified [1] as controlled vocabulary based and free text based systems at very high level. However, it is very difficult to create, maintain and scale a controlled vocabulary for CLIR systems in a general domain for a large corpus. Researchers came up with models that can be built on the full text of the corpus. The free text based system research can be broadly classified as corpus-based and knowledge-based aspects. Corpus-based systems may use parallel or comparable corpora, which are aligned at the word level, sentence level or passage level to learn models automatically. Knowledge-based systems might use bilingual dictionaries or ontologies, which form the handcrafted knowledge readily available for the systems to use. Hybrid systems were also built combining the knowledge-based and corpus-based approaches. Apart from these approaches, the extension of monolingual

IR techniques such as vector-based models, relevance modeling techniques [2] etc., to cross language IR were also explored.

In this work we have discussed our experiments on CLIR for Indian languages to English, where the queries are in Indian languages and the documents to be retrieved are in English. Experiments were carried out using queries in three Indian languages using the CLEF 2007 experimental setup. The three languages chosen were Bengali, Hindi and Telugu, which are predominantly spoken in the eastern, northern and southern parts of India, respectively.

2 Related Work

Very little work has been done in the past in the areas of IR and CLIR involving Indian languages. In the year 2003 a surprise language exercise [3] was conducted at ACM TALIP¹. 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. The International Institute of Information Technology (IIIT) in Hyderabad, India built a monolingual web search engine for various Indian languages, which is capable of retrieving information from multiple character encodings [6]. In the CLEF 2006 ad-hoc document retrieval task, Hindi and Telugu to English Cross Lingual Information Retrieval task [7] were reported by IIIT, Hyderabad.

Some research was previously carried out in the area of machine translation (MT) involving Indian languages [8], [9] etc. Most of the Indian language MT efforts involve studies on translating various Indian languages among themselves or translating contents from English to Indian language. Hence most of the Indian language resources available for the works are largely biased to these tasks. Recently, the Government of India has initiated a consortia project titled “Development of Cross-Lingual Information Access System” [10], where the query would be in any of the six different Indian languages (Bengali, Hindi, Marathi, Telugu, Tamil, Punjabi) and the output would be also in the language desired by the user.

3 Our Approach

The experiments carried out by us for CLEF 2007 are based on stemming, zonal indexing and TFIDF based ranking model with bilingual dictionary look up. There were no readily available bilingual dictionaries that could be used as databases for this work, so we had to develop bilingual dictionaries from the available resources in the Internet. The method of zonal indexing was applied on the English document collection after removing stop words and performing stemming operation. The keywords in the English document collection were indexed using the n-gram indexing methodology. The query terms were extracted from the topic files using bilingual dictionaries.

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

The Information Retrieval system was working on a TFIDF based ranking model. Query construction for the Telugu-English bilingual task was carried out by manually looking up the Telugu terms and choosing the translation from a human readable online Telugu-English dictionary. Machine-readable online bilingual dictionaries were used for the Bengali-English and Hindi-English bilingual tasks.

3.1 Zonal Indexing

In zonal indexing [11], a particular document is divided into n number of zones/regions, say, $w_1, w_2, \dots, w_n$. Then a weight is associated with each zone in such a way that the sum of all weights results in 1. Here, we divided each document into two zones, say, w_1 and w_2 . The zone ' w_1 ' contains the contents of the ED, PT, DK, EI, KH, HD and AU tags and ' w_2 ' region contains the contents of ID and TE tags of the Los Angeles Times (LA TIMES, 2002) documents. The weights heuristically assigned to w_1 and w_2 were 0.3 and 0.7, respectively. The contents of these two zones for all the documents were checked for stop words and then stemmed using the *Porter Stemmer* algorithm [12]. Relative term frequency of a content word in a document is then calculated in each of the w_1 and w_2 regions as the ratio of the number of occurrences of the content word in the region to the total number of content words present in that region. The relative term frequencies of any content word in the two regions are normalized and added together to get the relative term frequency of that content word in the entire document. These content words which could be multiwords were used as index keywords. The keywords in the English document collection were indexed using the n -gram indexing methodology. For every n -gram identified from the index keywords, all possible lower order $(n-1)$ grams starting from unigrams were considered as index keywords. The maximum value of n was considered to be 3. As an example, for the trigram "Indian Prime Minister" identified from an English document, the following were included in the index keywords:

Monograms: Indian, Prime, Minister

Bigrams: Indian Prime, Prime Minister

Trigram: Indian Prime Minister

The relevance of a document to an index keyword is calculated in terms of the product of the relative term frequency and inverse document frequency for the indexing keyword. We have considered the weighted sum of the relative term frequencies in each of the w_1 and w_2 zones. The inverse document frequency is calculated in terms of $\text{Log} [N/n_i]$ where, N is the number of documents in the collection and n_i is the number of documents in which the i th indexing keyword appears.

3.2 Query Construction

The title, description and narrative parts of the topic files are checked for stop words. Words/terms that appear in the topic files but are not significant as query words have been included in the stop word lists for each language. We have also prepared a list of words/terms for each language that identifies whether the query terms in the narrative parts provided with each topic talk about relevance/irrelevance of the query terms with respect to the topic. The narrative part of each topic in the topic files is checked

for these relevance/irrelevance stop words to mark the relevant or irrelevant keywords.

After the stop words have been deleted, the popular *Porter Stemming* [12] algorithm has been applied to remove the suffixes from the terms in the English topic file. For every n-gram, all possible n-1, n-2,...,1 grams (maximum value of n is 3) are extracted from the title, description and narration parts of the topic file. Two consecutive words are treated as the n-gram if no stop word appears in between them.

Indian languages are inflectional/agglutinative in nature and thus demand good stemming algorithms. Due to the absence of good stemmers for Indian languages, the words in the Bengali, Hindi and Telugu topic files are subjected to suffix stripping using manually prepared suffix lists in the respective languages. For every n-gram surrounded by stop words, all possible n-1, n-2, n-3,..., 1 grams (maximum value of n is 3) are extracted from the title, description and narrative parts of each English and Bengali topic and from the title part of each Hindi and Telugu topic. Only, n and 1 grams are extracted from the description and narrative parts of each Hindi topic file and from only the description part of the Telugu topic.

3.2.1 Query Translation

The available Bengali-English dictionary² was conveniently formatted for the machine-processing tasks. The Hindi-English dictionary was developed from the available English-Bengali and Bengali-Hindi machine-readable dictionaries. Initially, the English-Hindi dictionary was constructed. This dictionary was then converted into a Hindi-English dictionary for further use. A Telugu-English human readable online dictionary was used for query construction from Telugu topic files. Related works on dictionary construction can be found in [13].

The terms remaining after suffix removal are looked up in the corresponding bilingual Bengali/Hindi/Telugu to English dictionary. All English words/terms found in the bilingual dictionary for a word are considered, these may be synonyms or may correspond to different senses of the source language word. At present, we have not incorporated any technique to deal with the word sense disambiguation problem. Many of the terms may not be found in the bilingual dictionary, as the term may be a proper name or a word from a foreign language or a valid Indian language word, which did not occur in the dictionary. Dictionary look up may fail also in some cases due to the errors involved in the process of stemming and/or suffix removal. For handling dictionary look up failure cases, a transliteration from Indian languages to English was attempted assuming the word to be most likely a proper name not to be found in the bilingual dictionaries.

3.2.2 Query Transliteration

The transliteration engine is the modified joint source-channel model [14] based on the regular expression based alignment techniques. The Indian language NE is divided into Transliteration Units (TU) with patterns C^+M , where C represents a consonant or a vowel or a conjunct and M represents the vowel modifier or matra. An English NE is divided into TUs with patterns C^*V^* , where C represents a consonant and V represents a vowel. The system learns mappings automatically from the bilingual training corpus. The output of this mapping process is a decision-list classifier

² <http://dsal.uchicago.edu/dictionaries/biswas-bengali>

with collocated TUs in the source language and their equivalent TUs in collocation in the target language along with the probability of each decision obtained from the training corpus. Three different bilingual training sets namely, Bengali-English, Hindi-English and Telugu-English were developed to train the transliteration engine. The Bengali-English training set contains 25,000 manually created bilingual examples of proper names, particularly person and location names. The Hindi-English and Telugu-English bilingual training sets were developed from the Bengali-English training set. The Hindi-English and Telugu-English training sets contain 5,000 manually created bilingual training examples. The Indian language terms are thus translated and transliterated into the English terms accordingly. These translated/transliterated terms are then added together to form the English language query terms as part of query expansion.

3.3 Experiments

The evaluation document set, relevance judgments and the evaluation strategy are discussed in [15]. Three different runs were submitted related to the three Indian languages, one for each of the three languages, Bengali, Hindi and Telugu as our task in the ad-hoc bilingual track. Another run was submitted for English as a part of the ad-hoc monolingual task. Three runs were performed using the title, description and narration parts of the topic files for Bengali, Hindi and English. Only title and description parts of the topic file were considered for the bilingual Telugu-English run. All query words for a topic are searched in the index files. The *document id* and *relevance* are retrieved for each query term. Relevance of document to a query topic is calculated by the following method:

Relevance=(Sum of the relevance values for all relevant query terms - Sum of the relevance values for all irrelevant query terms).

3.4 CLEF 2007 Evaluation for Bengali-English, Hindi-English, Telugu-English Bilingual Ad-Hoc Task and English Monolingual Ad-Hoc Task

The run statistics for the four runs submitted to CLEF 2007 are described in Table 1. Clearly the geometric average precision metrics and its difference from mean average precision metrics suggests the lack of robustness in our system. There were certain topics that performed very well across the language pairs as well as for English also, but there were many topics where the performance was very low. The values of the evaluation metrics of Table 1 show that our system performs the best for the monolingual English task. As part of the bilingual ad-hoc tasks, the system performs best for the Telugu followed by Hindi and Bengali. The key to these higher values of the evaluation metrics in the Telugu-English bilingual run compared to other two bilingual runs (Hindi-English and Telugu-English) may be the manual tasks that were carried out during query processing phase. But it is also evident that the automatic runs for Hindi-English and Bengali-English tasks achieved a performance comparable to the manual run of Telugu-English. The overall relatively low performance of the system particularly with Indian language queries is indicative for the fact that simple techniques such as dictionary lookup with minimal lemmatization such as suffix removal may not be sufficient for the morphologically rich Indian languages CLIR. The relatively low performance of Bengali/Hindi emphasizes the need for broader

Table 1. Run Statistics

Run Id	MAP	R-Prec	GAP	B-Pref
AHB1L1BN2ENR1	10.18%	12.48%	2.81%	12.72%
AHB1L1HI2ENR1	10.86%	13.70%	2.78%	13.43%
AHB1L1TE2ENR1	11.28%	13.92%	2.76%	12.95%
AHMONOENR1	12.32%	14.40%	4.63%	13.68%

coverage of dictionaries and good morphological analyzer is inevitable for Bengali/Hindi CLIR in order to achieve a reasonable performance. Detailed statistics and plots of the bilingual runs can be found in the site, <http://10.2415/AH-B1L1-X2EN-CLEF2007.JADAVPUR.RunId>, where Run Id's are shown in Table 1. The monolingual English runs can be found in <http://10.2415/AH-B1L1-X2EN-CLEF2007.JADAVPUR.AHMONOENR1>.

3.5 Discussion

The unavailability of the appropriate machine-readable dictionaries often reduces the performance of our system. Absence of good stemmers for Indian languages is another big problem. Simple suffix removal may not be an ideal case always as it involves a lot of ambiguities. So, in order to deal with the highly inflective Indian languages we need robust stemmers. In addition, the terms remaining after suffix removal are looked up in the corresponding bilingual Bengali/Hindi/Telugu to English dictionary. All English words/terms found in the Bengali/Hindi/Telugu to English dictionary for a word are considered, these may be synonyms or may correspond to different senses of the source language word. Many of the terms may not be found in the bilingual dictionary, as the term is a proper name or a word from a foreign language or a valid Indian language word, which did not occur in the dictionary. We have used a transliteration model to deal with these dictionary look up failure cases. The accuracy of the transliteration system has a direct effect on the overall performance of the system. The use of a word sense disambiguation can be effective to improve the performance of the system.

4 Conclusion and Future Works

Our experiments suggest that simple TFIDF based ranking algorithms may not result in effective CLIR systems for Indian language queries. Any additional information added from corpora either resulting in source language query expansion or the target language query expansion or both could help. Machine-readable bilingual dictionaries with more coverage would have improved the results. An aligned bilingual parallel corpus would be an ideal resource to have in order to apply certain machine learning approaches. Application of word sense disambiguation methods on the translated query words would have a positive effect on the result. A robust stemmer is required for the highly inflective Indian languages. We would like to automate the query construction task of Telugu in future.

References

1. Oard, D.: Alternative Approaches for Cross Language Text Retrieval. In: AAAI Symposium on Cross Language Text and Speech Retrieval, USA (1997)
2. Lavrenko, V., Choquette, M., Croft, W.: Cross-Lingual Relevance Models. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM 2002, pp. 175–182. ACM Press, New York (2002)
3. Oard, D.: The Surprise Language Exercises. *ACM Transactions on Asian Language Information Processing* 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 2006: Proceedings of the 15th International Conference on World Wide Web, pp. 801–809 (2006)
7. Pingali, P., Varma, V.: Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006. In: Working Notes for the CLEF 2006 Workshop (Cross Language Adhoc Task), Alicante, Spain, 20–22 September (2006)
8. 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)
9. Naskar, S., Bandyopadhyay, S.: Use of Machine Translation in India: Current Status. In: Proceedings of MT SUMMIT-X, Phuket, Thailand, pp. 465–470 (2005)
10. CLIA Consortium: Cross Lingual Information Access System for Indian Languages. In: Demo/Exhibition of the 3rd International Joint Conference on Natural Language Processing, Hyderabad, India, pp. 973–975 (2008)
11. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, ch. 6. Cambridge University Press, Cambridge (2000)
12. Porter, M.F.: An Algorithm for Suffix Stripping. *Program* 14(3), 130–137 (1980)
13. Mayfield, J., McNamee, P.: Converting On-line Bilingual Dictionaries from Human-readable form to Machine-readable form. In: Proceedings of 25th Annual International ACM SIGIR Conference on Research and Development in Informational Retrieval, pp. 405–406. ACM Press, New York (2002)
14. Ekbal, A., Naskar, S., Bandyopadhyay, S.: A Modified Joint Source-Channel Model for Transliteration. In: Proceedings of the COLING/ACL, Sydney, Australia, pp. 191–198 (2006)
15. Nunzio, G.M.D., Ferro, N., Mandi, T., Peters, C.: CLEF 2007: Ad HOC Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)

Bengali and Hindi to English CLIR Evaluation

Debasis Mandal, Mayank Gupta, Sandipan Dandapat,
Pratyush Banerjee, and Sudeshna Sarkar

Department of Computer Science and Engineering
IIT Kharagpur, India - 721302
{debasis.mandal,pratyushb}@gmail.com,
{sandipan,mayank,sudeshna}@cse.iitkgp.ernet.in

Abstract. This paper presents a cross-language retrieval system for the retrieval of English documents in response to queries in Bengali and Hindi, as part of our participation in CLEF¹ 2007 Ad-hoc bilingual track. We followed the dictionary-based Machine Translation approach to generate the equivalent English query out of Indian language topics. Our main challenge was to work with a limited coverage dictionary (of coverage $\sim 20\%$) that was available for Hindi-English, and virtually non-existent dictionary for Bengali-English. So we depended mostly on a phonetic transliteration system to overcome this. The CLEF results point to the need for a rich bilingual lexicon, a translation disambiguator, Named Entity Recognizer and a better transliterator for CLIR involving Indian languages. The best MAP values for Bengali and Hindi CLIR for our experiment were 7.26% and 4.77%, which are 20% and 13% of our best monolingual retrieval, respectively.

1 Introduction

The growing number of multilingual web-accessible documents has benefitted many users who are able to read documents in more than one language. India, being a multilingual country of 22 official languages, has most of its inhabitants bilingual in nature, and exposed to English or Hindi (or both), in addition to their mother tongue. This necessitates the cross-language retrieval across the web, where the information need is expressed in a user's native language (the source language) and a ranked list of documents is returned in another language (target language). Since the language of query and documents to be retrieved are different, either the documents or queries need to be translated for CLIR. But the translation step tends to cause a reduction in the retrieval performance of CLIR as compared to monolingual information retrieval. Due to this reason, unambiguous translation is an important part of CLIR research. Various approaches involving parallel corpora, machine translation and bilingual dictionary have been experimented to address this problem [1,2,3,4]. However, in this paper we will restrict ourselves in the dictionary-based Machine Translation approach.

¹ Cross Language Evaluation Forum. <http://clef-campaign.org>

Bengali and Hindi are considered to be very resource-poor languages, in the sense that few language resources or tools (e.g. bilingual lexicon, morphological generator, parser etc) are available for them. This is due to the reason that much work has not yet been done in CLIR involving them. The other obstacle is the percentage of web contents for these languages, which is much less compared to other resource-rich languages, like English. Even within this limited content we faced several language-specific obstacles, like proprietary encodings of much of the web texts, that prohibited us to build the required training corpus for these languages. The scarcity of good parallel corpora restricted us to build the computational resources, like bilingual statistical lexicon and statistical transliterator. Moreover, the stemmers that were available for these languages usually make use of an extensive set of linguistic rules and thus lack comprehensive coverage. Furthermore, a named entity recognizer for Bengali and Hindi were also not available during the experiments [5].

Under this limited resource scenario, the sole objective of our participation in CLEF was to evaluate the basic CLIR system we had for Bengali and Hindi², and to explore the resource dependency, sources of improvement and comparability with other CLIR systems. This was our first participation in CLEF and we conducted six bilingual and three monolingual retrieval experiments for two language pairs: Hindi and Bengali to English. The investigation of CLEF evaluation results provided the necessary scope of improvement in the system and the importance of various IR components in great detail.

The rest of the paper is organized as follows. Section 2 presents some of the primitive and influencing works in CLIR involving Indian languages. The following section builds our CLIR model on the basis of bilingual lexicons, stemmers and transliterators. CLEF evaluations of our experiments and their analysis are presented in the subsequent section. We conclude this paper with a set of inferences and scope of future works.

2 Related Work

Cross-language retrieval involving Indian languages is relatively a new area of research among Natural Language Processing community, and first major work involving Hindi occurred only during TIDES Surprise Language exercise [7] in 2003. The objective of the exercise was to retrieve Hindi documents, provided by LDC (Linguistic Data Consortium), in response to queries in English. Interestingly, it was just an evolving field in India at that time and so no Indian university participated in the exercise. The five participants displayed a beautiful collaboration among them and submitted individual systems within one month period. They built a statistical lexicon out of parallel corpora [6,8,9] and used it to design Machine Translation based cross-lingual systems. The experiments had many interesting outcomes. Assigning TF.IDF weights on query terms and expanding query using training corpora were shown to improve the cross-language results even over

² Hindi and Bengali are world's fifth and seventh most spoken languages, respectively. Ethnologue: Languages of the World, 15th ed. (2005) <http://www.ethnologue.com>

Hindi monolingual runs [8]. Larkey et al. approached the problem using Language modeling approach [6] and showed the importance of a good stemmer for highly inflected languages, like Hindi. Finally, the exercise established the need of a good bilingual lexicon, query normalization, stop-words removal, stemming, query expansion with feedback and transliteration for the good result for Hindi.

The recent interest in cross-language research has given rise to a consortium for Cross-Language Information Access (CLIA) involving six Indian languages and premier research institutes across the country. As part of the ongoing research, several approaches have been tested and evaluated for CLIR in Indian languages in CLEF. The Language modeling coupled with Probabilistic transliteration, used by Larkey et al. [6] in surprise exercise, was also shown to be fruitful for Hindi and Telugu to English CLIR by Prasad et al. [10]. The approach also showed a significant improvement in performance over the simple dictionary-based Machine Translation. Manoj et al. performed Marathi and Hindi to English CLIR using Iterative Disambiguation Algorithm, which involves disambiguating multiple translations based on term-term co-occurrence statistics [11]. Jagadeesh et al. [12] had used a word alignment table, learned by a Statistical Machine Translation (SMT) system and trained on aligned parallel sentences, to convert the query into English. Sivaji et al. [13] has approached the problem for Hindi, Bengali and Telugu languages using a zonal-indexing approach on the corpus documents. In their approach, each document was first divided into some zones and then assigned some weights, the relative frequency of a term is then calculated based on zonal frequencies and thereafter used as an index keyword for query generation. Some of the other issues with the CLIA involving Indian languages and their feasible remedies are also discussed in [14,15,16].

3 Experiments

A basic dictionary-based Machine Translation approach, viz., tokenization, stop-words removal, stemming, bilingual dictionary look up and phonetic transliteration were followed to generate the equivalent English query out of Indian language topics. The main challenge of our experiment was to transliterate out-of-dictionary words properly and use limited bilingual lexicon efficiently. We had access to a Hindi-English bilingual lexicon³ of $\sim 26K$ Hindi words, a Bengali biochemical lexicon of $\sim 9K$ Bengali words, a Bengali morphological analyzer and a Hindi Stemmer. In order to achieve a successful retrieval under this limited resource, we adopted the following strategies: Structured Query Translations, phoneme-based followed by a list-based named entity transliterations, and performing no relevance judgment. Finally, the English query was fed into Lucene search engine and the documents were retrieved along with their normalized scores, which follows the Vector Space Model (VSM) of Information Retrieval. Lucene was also used for the tokenization and indexing of corpus documents.

³ ‘Shabdanjali’.

http://ltrc.iiit.net/onlineServices/Dictionaries/Dict_Frame.html

3.1 Structured Query Translation

After stemming of the topic words, the stemmed terms were looked up in the machine readable bilingual lexicon. If the term occurred in the dictionary, all of the corresponding translations were used to generate the final query. Parts-of-speech information of the topic words were not considered during translation. But many of those terms did not occur in the lexicon due to following reasons: limitations of the dictionary, improper stemming, the term is a foreign word or a named entity [10]. A close analysis showed that only 13.47% of the terms from ‘title+desc’ fields and 19.59% of the terms from ‘title+desc+narr’ fields were only found in the Hindi bilingual lexicon. For Bengali bilingual lexicon, the probability of finding a term dropped to below 5%.

3.2 Query Transliteration

The out-of-dictionary topic words were then transliterated into English using a phonetic transliteration system, assuming them to be *proper nouns*. The system works in the character level and converts every single Hindi or Bengali character in order to transliterate a word. But it produced multiple possibilities for every word, since English is not a phonetic language. For example, the Hindi term for *Australia* had four possible transliterations as output: *astreliya*, *astrelia*, *austreliya*, and *austrelia*. To disambiguate the transliterations, the terms were then matched against a manually-built named entity list with the help of an approximate string matching algorithm, *edit-distance algorithm*. The algorithm returns the best match of a term for pentagram statistics. For above example, the list correctly returned *Australia* as the final query term in cross-language runs.

Note that we did not expand the query using Pseudo Relevance Feedback (PRF) system. This is due to the fact that it sometimes does not improve the overall retrieval significantly for CLIR, rather hurts the performance by increasing noise towards the end retrievals [17]. Furthermore, it also increases the number of queries for which no relevant documents are returned, as shown in [8].

4 Results

The objective of Ad-Hoc Bilingual (X2EN) English task was to retrieve at least 1000 documents corresponding to each of the 50 queries from English target collection and submit them in ranked order. The data set and metrics for the Ad-Hoc track evaluation are described in detail in [18]. To evaluate the performance of our cross-language retrieval system, six bilingual runs were submitted for Bengali and Hindi, as shown in Table 1⁴. As a baseline, we also submitted three monolingual English runs consisting of various topic fields. For each of the Indian languages, the comparisons are made with respect to the best base run, viz., monolingual ‘title+desc’ run. The best values of Recall and MAP (Mean Average Precision) for the base run are 78.95% and 36.49%, respectively.

⁴ The DOI corresponding to a <Run ID> is <http://dx.doi.org/10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.<Run ID>>

Table 1. Cross-language runs submitted in CLEF 2007

Sl.#	Run ID	Topic Lang	Topic Field(s)
1	BENGALITITLE	Bengali	title
2	BENGALITITLEDESC	Bengali	title+desc
3	BENGALITITLEDESCNARR	Bengali	title+desc+narr
4	HINDITITLE	Hindi	title
5	HINDITITLEDESC	Hindi	title+desc
6	HINDITITLEDESCNARR	Hindi	title+desc+narr

The results of our cross-language task are summarized in Table 2 and Table 3. Table 2 shows that the recall gradually improved with the addition of more relevant terms from the topic fields for Bengali, as expected, but the same did not repeat for Hindi. This result was a surprise to us as we had used a bilingual lexicon of superior performance for Hindi. A careful analysis revealed that the value of MAP is also poorer for Hindi, as seen in Table 3, contrary to our expectation. Moreover, variations in the values of MAP and R-precision over different topic fields are not much for Hindi, as compared to Bengali. However, the precision values with respect to top 5, 10 and 20 retrievals demonstrate a steady increase for each of them.

Table 2. Summary of bilingual runs of the Experiment

Run ID	Relevant Docs	Relevant Retrieved	Recall (in %)	% mono	B-Pref
BENGALITITLE	2247	608	27.60	34.96	5.43
BENGALITITLEDESC	2247	851	37.87	47.97	10.38
BENGALITITLEDESCNARR	2247	906	40.32	51.07	11.21
HINDITITLE	2247	708	31.51	39.91	9.95
HINDITITLEDESC	2247	687	30.57	38.72	11.58
HINDITITLEDESCNARR	2247	696	30.97	39.23	12.02

The anomalous behavior of Hindi can be explained in terms of translation disambiguation during query generation. Query wise score breakup revealed that the queries with more named entities always provided better results than their counterparts. With the increase of lexical entries and Structured Query Translation (SQT), more and more ‘noisy words’ were incorporated into final query in the absence of any translation disambiguation algorithm, thus bringing down the overall performance. The average English translations per Hindi word in the lexicon were 1.29, with 14.89% Hindi words having two or more translations. For example, the Hindi word ‘rokanA’ (to stop) had 20 translations, making it highly susceptible towards noise. Figure 1 shows the frequency distribution of dictionary entries with their corresponding number of translations in the Hindi

bilingual dictionary. It is also evident from Table 3 that adding extra information to query through ‘desc’ field increases the performance of the system, but adding ‘narr’ field has not improved the result significantly. The post-CLEF analysis revealed that this field constituted two parts: relevance and irrelevance, and was meant to prune out the irrelevant documents during retrieval. But we did not make any effort in preventing the irrelevant retrieval in our IR model.

Table 3. Precision results (in %) for bilingual runs in CLEF 2007

Run ID	MAP	% mono	R-Prec	P@5	P@10	P@20
BENGALITITLE	4.98	13.65	5.86	4.80	6.60	7.00
BENGALITITLEDESC	7.26	20.00	8.53	10.00	10.20	8.80
BENGALITITLEDESCNARR	7.19	19.70	9.00	11.60	10.80	10.70
HINDITITLE	4.77	13.07	5.34	8.40	6.40	5.40
HINDITITLEDESC	4.39	12.03	5.19	9.20	8.60	7.10
HINDITITLEDESCNARR	4.77	13.07	5.76	10.40	8.40	7.30

The results in Table 3 show that the best MAP values for Bengali and Hindi CLIR for our experiment are 7.26% and 4.77% which are 20% and 13% of our best base run, respectively. Although the result of Bengali is comparable (10.18%) with only other participant for the language in CLEF 2007 [13], the results for Hindi in our experiment was much poorer than the best entry (29.52%) [11]. Lack of a good bilingual lexicon can be attributed as the primary reason for our poor result.

The other shortcoming of our system is the homogeneous distribution of precision with respect to retrieved documents and interpolated recall, as evident from Figure 2. This clearly demands for a good feedback system (e.g. Pseudo Relevance Feedback) to push the most relevant documents to the top. Apart from the costly query refinement operation, improvement can also be made by

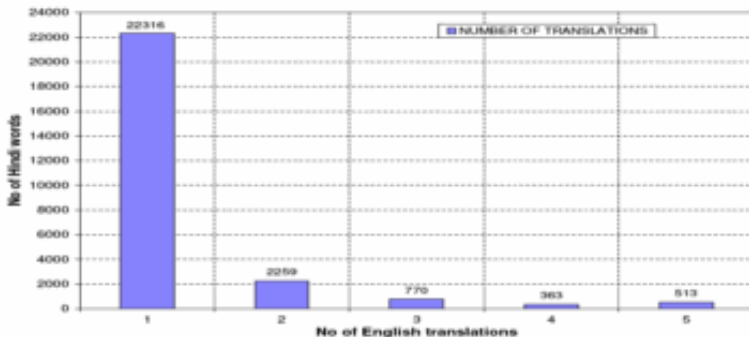


Fig. 1. Frequency distribution of number of translations in Hindi bilingual dictionary

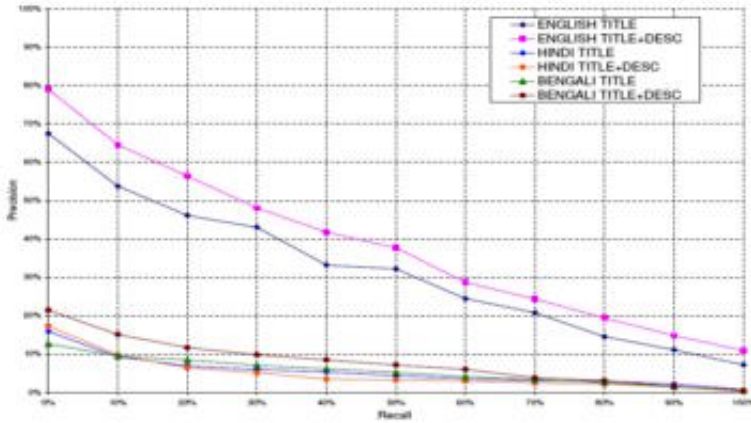


Fig. 2. Recall vs. Precision results for the Experiment

identifying the named entities in the query and assigning them a higher relative weight with respect to other query terms.

5 Conclusions and Future Works

This paper described an experiment of Bengali and Hindi to English cross-language text retrieval as part of CLEF 2007, its evaluation results and few post-evaluation analyses. The poorer performance of our system with respect to other resource-rich participants clearly pointed out the necessity of a rich bilingual lexicon, a good transliteration system, and a relevance feedback system. Further, part of speech (POS) information will help to disambiguate the translations. Performance of the stemmer also has an important role in cross-language retrieval for morphologically rich languages, like Bengali and Hindi.

Our future work includes building named entity recognizers and efficient transliteration system based on statistical and linguistic rules. We would also like to analyze the effect of feedback system in cross-language query expansion. Language modeling is another approach we would like to test upon for a better cross-language retrieval involving Indian languages.

Acknowledgment

We would like to thank Mr. Sunandan Chakraborty of the Department of Computer Science & Engineering, IIT Kharagpur, for his generous help to resolve various programming issues during integration of the system.

References

1. Hull, D., Grefenstette, G.: Querying across languages: A dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 49–57 (1996)
2. Diekema, A.R.: Translation Events in Cross-Language Information Retrieval. ACM SIGIR Forum 38(1) (2004)
3. Bertoldi, N., Federico, M.: Statistical Models for Monolingual and Bilingual Information Retrieval. Information Retrieval 7, 53–72 (2004)
4. Monz, C., Dorr, B.: Iterative Translation Disambiguation for Cross-Language Information Retrieval. In: SIGIR 2005, Salvador, Brazil, pp. 520–527 (2005)
5. Mandal, D., Dandapat, S., Gupta, M., Banerjee, P., Sarkar, S.: Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
6. 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)
7. Oard, D.W.: The surprise language exercises. ACM Transactions on Asian Language Information Processing (TALIP) 2(2), 79–84 (2003)
8. Xu, J., Weischedel, R.: Cross-Lingual Retrieval for Hindi. ACM Transactions on Asian Language Information Processing (TALIP) 2(1), 164–168 (2003)
9. Allan, J., Lavrenko, V., Connell, M.E.: A Month to Topic Detection and Tracking in Hindi. ACM Transactions on Asian Language Processing (TALIP) 2(2), 85–100 (2003)
10. Pingali, P., Tune, K.K., Varma, V.: Hindi, Telugu, Oromo, English CLIR Evaluation. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
11. Chinnakotla, M.K., Ranadive, S., Bhattacharyya, P., Damani, O.P.: Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
12. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
13. Bandyopadhyay, S., Mondal, T., Naskar, S.K., Ekbal, A., Haque, R., Godavarthy, S.R.: Bengali, Hindi and Telugu to English Ad-hoc Bilingual task at CLEF 2007. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
14. Pingali, P., Jagarlamudi, J., Varma, V.: Webkhoz: Indian language IR from Multiple Character Encodings. In: International World Wide Web Conference (2006)
15. Pingali, P., Varma, V.: IIIT Hyderabad at CLEF 2007 Adhoc Indian Language CLIR task. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
16. Pingali, P., Varma, V.: Multilingual Indexing Support for CLIR using Language Modeling. In: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering (2007)
17. Clough, P., Sanderson, M.: Measuring Pseudo Relevance Feedback & CLIR. In: SIGIR 2004, UK (2004)
18. Nunzio, G.M.D., Ferro, N., Mandl, T., Peters, C.: CLEF 2007: Ad-Hoc Track Overview. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)

Improving Recall for Hindi, Telugu, Oromo to English CLIR

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 the Language Technologies Research Centre (LTRC, IIIT-Hyderabad) as part of our participation in the ad-hoc track of CLEF 2007. We present approaches to improve recall of query translation by handling morphological and spelling variations in source language keywords. We also present experiments using query expansion in CLIR using a source language monolingual corpus for Hindi, Telugu and English. We also present the effect of using an Oromo stemmer in Oromo-English CLIR system and report results using the CLEF 2007 dataset.

1 Introduction

There is a growing interest in CLIR research involving languages with little or no language resources. In this paper, we discuss our CLIR experiments using the CLEF 2007 dataset for Hindi-English, Telugu-English and Oromo-English language pairs. We first present an overview of previous research work related to CLIR involving these languages. Task focused workshops conducted at CLEF 2006 and CLEF 2007 have proved very helpful in evaluating CLIR systems for the above mentioned languages. For instance, techniques of using existing bilingual lexica were tried for Hindi and Telugu to English CLIR in [1] in CLEF 2006. Similarly, Pingali et. al., [1] attempted to apply query expansion techniques and statistical transliteration of source language out-of-vocabulary (OOV) words to overcome the recall problem of dictionary based keyword translation. Jagarlamudi and Kumaran [2] used parallel corpora while Chinnakotla et. al., [3] used a bilingual lexicon for Hindi English CLIR and achieved good performance. In CLEF 2007 [4] we attempted to build bilingual dictionaries using topical similarity by choosing vocabulary from a web search engine index [5] and demonstrated that such dictionaries perform very well even with fewer entries. Apart from these recent experiments in Indian language to English CLIR tasks, previously, a surprise language exercise [6] was conducted at ACM TALIP¹ in 2003. This task

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

was quite different from CLEF 2006 and 2007 tasks since the source language in this task was English. The ACM TALIP task provided some insights into the issues involved in processing Indian language content.

Likewise, very little CLIR research was done using African indigenous languages including major Ethiopian languages. In the recent past, we conducted CLIR experiments [1, 7] using Afaan Oromo. A case study for Zulu (one of the major languages in South Africa) was reported by [8] in relation to application for cross-lingual information access to knowledge databases. Another similar study was undertaken by Cosijn et. al., [9] on Afrikaans-English cross-language information retrieval. More recently, different dictionary-based Amharic-English and Amharic-French CLIR experiments were conducted at a series of CLEF ad hoc tracks [10, 11]. In the following sections we discuss the CLIR experiments we conducted using the CLEF 2007 dataset.

2 Hindi, Telugu, Oromo to English CLIR Experiments

Our submission to CLEF 2007 uses a vector based ranking model with a bilingual lexicon using word translations. Out of vocabulary words or OOVs for Indian language task are handled using a probabilistic algorithm as mentioned in Pingali et. al., [1]. For Oromo queries, we handle them using a named entity dictionary. Document retrieval is achieved using an extended boolean model where queries are constructed using the Boolean operators among keywords and the occurrence of keywords in various types of metadata is given a different weight. The various fields that were used while retrieving the documents are mentioned in section 2.1. The ranking is achieved using a vector based ranking model using a variant of TFIDF ranking algorithm. We used the Lucene framework to index the English documents. All the English documents were stemmed and stop words were eliminated to obtain the index terms. These terms were indexed using the Lucene² search engine using the TFIDF similarity metric.

2.1 Query Translation and Formulation

For a CLIR task, query translation need not be a true translation of the given source language query. In other words, the target language output produced need not be well formed and human readable and the only goal of such a translation is to obtain a topically similar translation in the target language to enable proper retrieval. In our system, a given source language query is translated using word by word translation. For Indian language tasks, each source language word is looked up in the bilingual dictionaries for exact match as well as all words having the same prefix as the given source language word beyond a fixed threshold. If a given source language word does not occur in the bilingual dictionary, one character at a time is removed from the end of the word until a matching word with same prefix is found in the dictionary. This heuristic for dictionary lookup helps in translating source language words even if their morphological variants or

² <http://lucene.apache.org>

compound words are present in the dictionary. We used a Hindi-English dictionary with 5,175 entries and Telugu-English dictionary with 26,182 entries which are discussed in [4]. For Oromo query translations, we apply an Oromo light stemmer [12] on both the queries and the dictionary words before looking up the meanings.

Traditionally IR systems use the notion of fields to treat different types of metadata related to a document with different weights. Queries are constructed to search for keywords and weigh them using some prior weights assigned intuitively to a given metadata type. For example, the title of a document can be viewed as a metadata of the given article and a keyword found in the title might be deemed to be more important than one found in the document's body. In our system, we used three such fields for each document. Two of the fields *title*, *body* were explicitly provided in the given corpus, while we derived a new field called *summary* of a document and chose the first 50 words of the document body as a *summary*. Different boost scores were given to each of these fields such that *title* of the document is deemed most important followed by *summary*, then followed by the *body* of the document. In order to provide different weights to terms based on fields, a combination of term and the field name is treated as a unique entry in the inverted index. In other words, a given term occurring in both title and body are treated as two different terms and their term frequencies and document frequencies are computed individually.

As mentioned in the beginning of this section, our retrieval approach is to translate the given source language query keywords using a bilingual dictionary. We translate one term at a time and do not handle multi-word expressions and phrases in queries. However, the algorithm itself does not require any modification to be able to handle multi-term queries. It would suffice if the multi-word expressions along with their meanings are also stored in the same dictionary as the single word dictionary. Our lookup algorithm first tries to lookup entries containing longest source language expressions. This is achieved since the lookup program also internally represents the dictionary using an inverted index data structure.

Once the source language queries are translated and transliterated, the resultant English keywords used to construct boolean queries using boolean AND and OR operators. Assume the index model to contain the set of fields/metadata as $F = f_1, f_2 \dots f_m$ and the source language query $S = s_1, s_2, \dots, s_n$, and every source language keyword s_i results in multiple target language keywords. Let t_{ij} be the j^{th} translation of source language keyword s_i . In our experiments we primarily construct a disjunctive type query Q_{disj} and a hybrid query Q_{hyb} for every k^{th} field from F as

$$Q_{disj,k} = w_k \cdot \bigcup_{i,j} t_{ij} \quad (1)$$

$$Q_{hyb,k} = w_k \cdot \bigcap_i \bigcup_j t_{ij} \quad (2)$$

where w_k is the boost weight given to the k^{th} field. Finally the multiple field queries are again combined using a boolean OR operator. We report various runs

based on the boolean operations on the queries and the fields on which retrieval is performed in the evaluation section.

It is evident from our approach that we do not make any efforts to identify the irrelevant documents in the search process. For this reason we did not use the *narrative* information in the topics for any of our runs. For Oromo-English CLIR tasks we experimented with *title*, *description* and *narrative* fields of the given topics in various runs.

2.2 Query Expansion

In this paper, we also present some experiments by expanding the given queries using a monolingual corpus in Hindi, Telugu and English. The *title* from the given query is used to initially retrieve a set of documents from a monolingual web search engine [5] which contains about 1 million Hindi, 200,000 Telugu and 1 million English documents respectively. A set of text fragments with a fixed window length containing keywords in context of the given *title* keywords is enumerated. A TFIDF model of these keywords in context is then built assuming the combination of all such text fragments to be a single document in the given monolingual corpus. A list of top ' N ' words from these words are used to expand the original input query title along with its description field. We report experiments on the role of query expansion in CLIR in the next section. We present a few examples of expanded query keywords in the Table 1. It can be observed from the Table that the generated keywords seem to topically correlate to the original input query and hence may be very useful to improve recall in CLIR task. We also tested its effect on monolingual task and found it to slightly improve recall.

Table 1. Example of additional keywords added to query using query expansion from top 20 documents

Query Keywords	Expanded Keywords
వోలీసుగా నడించడం	వోలీసాఫీసర్, పాత్రలో, అఫీసర్, కమిషనర్, అధికారి, అవా- ర్డు, పీరోపీరోయిన్లుగా, పీరోగా, పీరోయిన్, చిత్రం, వోకిరి, పాత్రలో, శంకర్, దర్శకత్వంలో, తదితరులు, సినిమాలు
మూరో ద్రవ్యోల్బణం	వారాంతంలో, శాతం, తగ్గింది, అంతర్జాతీయ, మార్కెట్, త- గ్గిన, ఇంతటిస్థాయిలో, పెరిగింది, పెరగడమే, ధరలు, మూ- రోపు, మూరోపిమన్, మూసిమన్, డాలర్, ఎలువ
यूरो की कीमत में वृद्धि	डालर जर्मनी अध्ययन मांग खरीदी फीसदी बाजार अमरीकी आर्थिक

3 Experiments and Discussion

Two runs were submitted related to the Indian languages, one with Hindi queries and one with Telugu queries. For the Oromo task, 3 runs were submitted during

Table 2. Run Descriptions

Run ID	Language Pair	Description
MONO	English	Monolingual run, using title, body and summary fields. Title keywords are combined using hybrid query as described in the previous section. CLEF official submission.
MDISJ	English	Monolingual run, using title, body and summary fields. All keywords are combined using boolean OR operator.
EEXP	English	Only body text is used. Source language keywords expanded and combined using boolean OR.
HNOSUM	Hindi - English	Title and body fields are used. Title keywords are combined using boolean AND.
HITD	Hindi - English	Uses title, body and summary fields. Title keywords are combined using boolean AND across translations. CLEF official submission.
HDISJ	Hindi - English	Only body text is used. All translated keywords combined using boolean OR.
HEXP	Hindi - English	Only body text is used. Source language keywords expanded, translated and combined using boolean OR.
TNOSUM	Telugu - English	Title and body fields are used. Title keywords are combined using boolean AND.
TETD	Telugu - English	Uses title, body and summary fields. Title keywords are combined using boolean AND across translations. CLEF official submission.
TDISJ	Telugu - English	Only body text is used. All translated keywords combined using boolean OR.
TEXP	Telugu - English	Only body text is used. Source language keywords expanded, translated and combined using boolean OR.
NOST_OMT07	Oromo-English	Only body text is used. Source language title keywords translated without stemming and combined using boolean OR .
NOST_OMTD07	Oromo-English	Only body text is used. Source language title, desc keywords translated without stemming and combined using boolean OR .
NOST_OMTDN07	Oromo-English	Only body text is used. Source language title, desc, narr keywords translated without stemming and combined using boolean OR .
OMT07	Oromo-English	Only body text is used. Source language title keywords are stemmed, translated and combined using boolean OR .
OMTD07	Oromo-English	Only body text is used. Source language title, desc keywords are stemmed, translated and combined using boolean OR .
OMTDN07	Oromo-English	Only body text is used. Source language title, desc, narr keywords are stemmed, translated and combined using boolean OR .

the CLEF task. A monolingual run was also submitted to obtain a baseline performance. After CLEF released the relevance judgements we conducted some more experiments. Table 3 describes the various runs we report in this section.

The average metrics for each of the runs mentioned in Table 3 are described in Table 3. The first column in Table 3 mentions the RUNID and the remaining of each column represents the various metrics. Each row gives a comparison of a given metric across all the runs. The metrics listed are as provided by the TREC evaluation package³. Of these metrics, we find *rel_{ret}*, *map*, *bpref* and *P10* values, which are relevant documents retrieved, mean average precision, binary preference and precision for first 10 results, to be interesting. Apart from these metrics we also report the number of relevant documents (*rel*) and mean reciprocal rank (*r_{-rank}*) metrics. From the run statistics it can be observed that

³ TREC provides a *trec_eval* package for evaluating IR systems.

the Hindi-English CLIR performs reasonably well even when the dictionary is very small around 5,000 words. Also from *P10*, it can be observed that systems using boolean AND operator with appropriate boosting of metadata results in better ranking. However, such systems result in lower recall. It can be observed from *rel_{ret}* of disjunctive runs MDISJ, TDISJ and HDISJ that the system is able to retrieve more relevant documents when queries are combined using boolean OR operator. It can also be observed that using a summary as a metadata in retrieval might help when the translation quality is low. This fact can be observed from better performance of HNOSUM run for Hindi, which performs better than HITD. However, use of a summary results in lower performance when the translation quality is better, which can be observed from TETD and TNOSUM. The role of monolingual query expansion applied to Hindi-English and Telugu-English CLIR tasks can be observed from HEXP and TEXP runs. It can be observed from *rel_{ret}* that source language query expansion improves the recall of CLIR. When a query expansion technique is applied to English monolingual retrieval, the improvement in performance is not significant which can be observed from the EEXP run. However, in the case of CLIR, we found the recall to increase a lot since query expansion increases the probability of translating a source language keyword into target language even when the keyword is not found in the bilingual lexicon. For example, Table 1 shows the keywords added for the given Telugu and Hindi queries. It can be observed that the added additional keywords are highly topically similar to the original query keyword. Therefore, even if the original query keyword did not have an entry in the dictionary, at least a few keywords translated from the expanded query seem to improve recall in HEXP and TEXP runs.

Table 3. Run Statistics for 50 queries of CLEF 2007 dataset

RUNID	ret	rel	rel _{ret}	map	gm _{ap}	R-prec	bpref	r _{rank}	P10
MDISJ	50000	2247	1952	0.4003	0.3335	0.4084	0.3980	0.7161	0.4920
MONO	50000	2247	1629	0.3687	0.2526	0.3979	0.3850	0.6584	0.4520
EEXP	50000	2247	1982	0.4115	0.3332	0.4025	0.4057	0.7671	0.4920
TETD	50000	2247	1275	0.2155	0.0834	0.2467	0.2868	0.5160	0.3060
TETDNOSUM	50000	2247	1456	0.2370	0.0950	0.2702	0.3083	0.4814	0.3060
TETDDISJ	50000	2247	1517	0.2170	0.0993	0.2478	0.2750	0.5419	0.3080
TEXP	50000	2247	1444	0.2648	0.1507	0.3029	0.2987	0.6100	0.3760
HITD	50000	2247	958	0.1560	0.0319	0.1689	0.2104	0.3778	0.1820
HITDNOSUM	50000	2247	1132	0.1432	0.0321	0.1557	0.2026	0.3270	0.2000
HITDDISJ	50000	2247	1123	0.1331	0.0321	0.1566	0.2005	0.3416	0.1960
HEXP	50000	2247	1373	0.1821	0.0430	0.1993	0.2247	0.4205	0.2520
NOST_OMT07	50000	2247	1224	0.1736	0.0374	0.1808	0.2226	0.3663	0.2304
NOST_OMTD07	50000	2247	1333	0.2010	0.0619	0.1955	0.2349	0.4253	0.2938
NOST_OMTDN07	50000	2247	1439	0.2038	0.0619	0.2217	0.2524	0.4454	0.3140
OMT07	50000	2247	1554	0.2420	0.1427	0.2624	0.2635	0.6032	0.3380
OMTD07	50000	2247	1707	0.2991	0.2020	0.3063	0.3027	0.7037	0.4200
OMTDN07	50000	2247	1693	0.2894	0.1967	0.2973	0.2987	0.6852	0.4320

In our experiments with Oromo queries, the top performance was achieved by our stemmed title and description run (OMTD07, MAP of 0.2991) and is about 67.95% of the best official CLEF 2007 English monolingual baseline (which were achieved by other participants at CLEF-2007). We feel our current achievements are quite significant and encouraging results given the very limited linguistic resources that we have employed in our Oromo-English retrieval experiments.

Comparing the Oromo-English runs with the Indian language runs we notice that the Oromo-English runs perform significantly better than the Hindi-English and Telugu-English runs. The better performance can be attributed to the quality and size of the dictionary used for Oromo-English when compared with Hindi-English. While dictionary was bigger in the case of Telugu-English runs, we observe that the characteristics of Telugu when compared with Oromo are quite different in the aspects of morphology. Telugu words can be highly inflectional to the extent that complete sentences can be written as single words which is not the case with Oromo. Such a difference demands for a larger dictionary and a more sophisticated word segmentation program for Telugu.

4 Conclusion

Our experiments using CLEF 2007 data suggest that the performance of a CLIR system heavily depends on the type and quality of the resources being used. While the underlying IR model combined with some additional query enriching techniques such as query expansion is also important and can play a role in the quality of a CLIR output, we found the main contribution to performance coming from the ability to convert a source language information need into the target language. We showed that, by using simple techniques to quickly create dictionaries, one can maximize the probability of retrieving the relevant documents. This was evident from the fact that our Hindi-English CLIR system used a very small dictionary of the size of 5,175 words, many of them containing variants of same words, implying an even small number of unique root word dictionary. However, the reason for success of this resource in a CLIR task is that, the choice of source language words in the dictionary is motivated by the TFIDF measure of the words from a sufficiently large corpus. Moreover the dictionary creators keyed-in meanings with an IR application in mind, instead of attempting to create an exact synonym dictionary. Also, the restriction on the number of keywords one can type for a given source language word enabled us to capture the homonyms instead of many polysemous variants. This also shows that one might be better off with task specific or task tailored dictionaries since different applications have different constraints and levels of tolerance to error.

The results we have obtained this year in our official and unofficial Oromo-English retrieval experiments show significant improvement over the CLEF 2006 experiments. We have tested and analyzed the impacts of an Afaan Oromo light stemmer on the overall performances of our Oromo-English CLIR system. The application of our light stemmer has significantly improved the performances of our CLIR system in all of our current experiments. Our analysis of errors

in performance are directly related to the coverage of our bilingual lexicon and hence believe that increasing the size of dictionary would directly improve the performance of our system.

References

1. Peters, C., et al. (eds.): CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
2. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 80–87. Springer, Heidelberg (2008)
3. Chinnakotla, M.K., Ranadive, S., Bhattacharyya, P., Damani, O.P.: Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 111–118. Springer, Heidelberg (2008)
4. Pingali, P., Varma, V.: IIIT Hyderabad at CLEF 2007 - Adhoc Indian Language CLIR task. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
5. Pingali, P., Jagarlamudi, J., Varma, V.: Webkhoj: Indian language ir from multiple character encodings. In: WWW 2006: Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, pp. 801–809. ACM Press, New York (2006)
6. Oard, D.W.: The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)* 2(2), 79–84 (2003)
7. Pingali, P., Varma, V., Tune, K.K.: Evaluation of Oromo-English Cross-Language Information Retrieval. In: *IJCAI 2007 Workshop on CLIA*, Hyderabad, India (2007)
8. 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)
9. Cosijn, E., Keskustalo, H., Pirkola, A.: Afrikaans - English Cross-language Information Retrieval. In: *Proceedings of the 3rd biennial DISSAnet Conference*, Pretoria (2004)
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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
12. Tune, K.K., Varma, V.: Oromo-English Information Retrieval Experiments at CLEF 2007 (2007)

Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation

Manoj Kumar Chinnakotla*, Sagar Ranadive, Om P. Damani,
and Pushpak Bhattacharyya

Department of Computer Science and Engineering, IIT Bombay, India
{manoj,sagar,damani,pb}@cse.iitb.ac.in

Abstract. In this paper, we present our Hindi to English and Marathi to English CLIR systems developed as part of our participation in the CLEF 2007 Ad-Hoc Bilingual task. We take a query translation based approach using bi-lingual dictionaries. Query words not found in the dictionary are transliterated using a simple rule based transliteration approach. The resultant transliteration is then compared with the unique words of the corpus to return the ‘k’ words most similar to the transliterated word. The resulting multiple translation/transliteration choices for each query word are disambiguated using an iterative page-rank style algorithm which, based on term-term co-occurrence statistics, produces the final translated query. Using the above approach, for Hindi, we achieve a Mean Average Precision (MAP) of 0.2366 using title and a MAP of 0.2952 using title and description. For Marathi, we achieve a MAP of 0.2163 using title.

1 Introduction

The World Wide Web (WWW), a rich source of information, is growing at an enormous rate. Although English still remains the dominant language on the Web, global internet usage statistics reveal that the number of non-English internet users is steadily on the rise. Hence, making this huge repository of information, which is available in English, accessible to non-English users worldwide is an important challenge in recent times.

Cross-Lingual Information Retrieval (CLIR) systems allow the users to pose the query in a language (*source language*) which is different from the language (*target language*) of the documents that are searched. This enables users to express their information need in their native language while the CLIR system takes care of matching it appropriately with the relevant documents in the target language. To help the user in the identification of relevant documents, each result in the final ranked list of documents is usually accompanied by an automatically generated short summary snippet in the source language. Using this, the user could single out the relevant documents for complete translation into the source language.

* Supported by a fellowship award from Infosys Technologies.

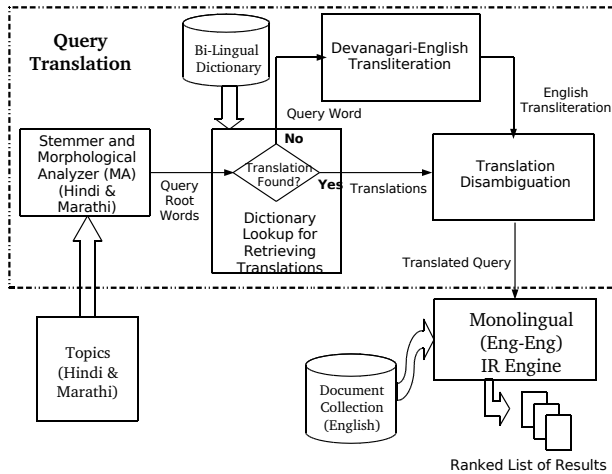


Fig. 1. CLIR System Architecture

Hindi is the official language of India along with English and according to *Ethnologue*¹, it is the fifth most spoken language in the world. *Marathi* is a widely spoken language in the state of Maharashtra. Both Hindi and Marathi use the “Devanagari” script.

In this paper, we describe our Hindi to English and Marathi to English CLIR approaches for the CLEF 2007 Ad-Hoc Bilingual task. The architecture of our CLIR system is shown in Figure 1. We use a *Query Translation* based approach in our system since it is efficient to translate the query vis-a-vis documents. It also offers the flexibility of adding cross-lingual capability to an existing monolingual IR engine by just adding the query translation module. We use machine-readable bi-lingual Hindi to English and Marathi to English dictionaries created by Center for Indian Language Technologies (CFILT), IIT Bombay for query translation. The Hindi to English bi-lingual dictionary has around 115,571 entries and is also available online². The Marathi to English bi-lingual has less coverage and has around 6110 entries.

Hindi and Marathi, like other Indian languages, are morphologically rich. Therefore, we stem the query words before looking up their entries in the bi-lingual dictionary. In case of a match, all possible translations from the dictionary are returned. In case a match is not found, the word is transliterated by the Devanagari to English transliteration module. The above module, based on a simple lookup table and index, returns top three English words from the corpus which are most similar to the source query word. Finally, the translation disambiguation module disambiguates the multiple translations/transliterations returned for the query and returns the most probable English translation of the

¹ <http://www.ethnologue.com>

² http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/dict_search_user.php

Table 1. A sample CLEF 2007 Hindi Topic: Number 445

<pre> <num>10.2452/445-AH</num> <title>प्रिन्स हैरी और नशीली दवाएं</title> <desc>ऐसे दस्तवेज खोजिये जिनमे प्रिन्स हैरी द्वारा नशीली दवाएं ग्रहण किए जाने की कोई रिपोर्ट हो</desc> </pre>

original query. The translated query is fired against the monolingual IR engine to retrieve the final ranked list of documents as results.

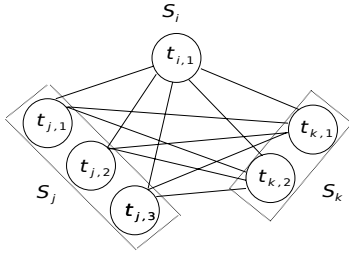
The organization of the paper is as follows: Section 2 presents the approach used for *Query Transliteration*. Section 3 explains the *Translation Disambiguation* module. Section 4 describes the experimental setup, discusses the results and also presents the error analysis. Finally, Section 5 concludes the paper highlighting some potential directions for future work.

2 Devanagari to English Transliteration

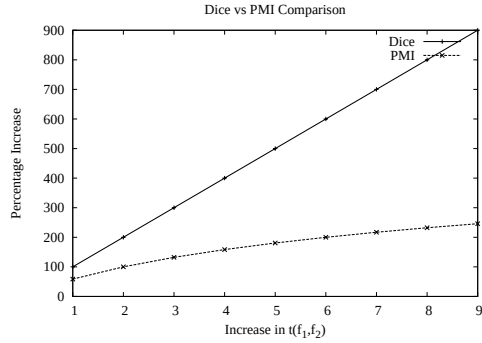
Many words of English origin like names of people, places and organizations, are likely to be used as part of the Hindi or Marathi query. Such words are usually not found in the Hindi to English and Marathi to English bi-lingual dictionaries. Table 1 presents an example Hindi topic from CLEF 2007. In the above topic, the word प्रिन्स हैरी is *Prince Harry* written in Devanagari. Such words need to be *transliterated* into English. We use a simple rule based approach which utilizes the corpus to identify the closest possible transliterations for a given Hindi/Marathi word.

We create a lookup table which gives the roman letter transliteration for each Devanagari letter. Since English is not a phonetic language, multiple transliterations are possible for each Devanagari letter. In our current work, we only use a single transliteration for each Devanagari letter. The English transliteration is produced by scanning a Devanagari word from left to right replacing each letter with its corresponding entry from the lookup table. The above approach produces many transliterations which are not valid English words. For example, for the word आस्ट्रेलियाई (Australian), the transliteration based on the above approach will be *astreliyai* which is not a valid word in English. Hence, instead of directly using the transliteration output, we compare it with the indexed words in the corpus and choose the ‘k’ most similar indexed words in terms of *string edit distance*. For computing the string edit distance, we use the dynamic programming based implementation of *Levenshtein Distance* [1] metric.

Using the above technique, the top 3 closest transliterations for आस्ट्रेलियाई were *australian*, *australia* and *estrella*. Note that we pick the top 3 choices even if our preliminary transliteration is a valid English word. The final choice of transliteration for the source term is made by the translation disambiguation module based on the term-term co-occurrence statistics of the transliteration with translations/transliterations of other query terms.



(a) Co-occurrence Graph [2]



(b) Dice vs. PMI

Fig. 2. Translation Disambiguation: Co-occurrence Graph for Disambiguating Translations and Transliterations, Comparison of Dice Coefficient and PMI

3 Translation Disambiguation

Given the various translation and transliteration choices for the query, the Translation Disambiguation module, out of the various possible combinations, selects the *most probable* translation of the input query Q . The context within a query, although small, provides important clues for choosing the right translations/transliterations of a given query word. For example, for a query “नदी जल” (River Water), the translation for नदी is $\{river\}$ and the translations for जल are $\{water, to\ burn\}$. Here, based on the context, we can see that the choice of translation for the second word is *water* since the combination $\{river, water\}$ is more likely to co-occur in the corpus than $\{river, burn\}$.

Consider a query with three words $Q = \{s_i, s_j, s_k\}$. Let $tr(s_j) = \{t_{j,1}, t_{j,2}, \dots, t_{j,l}\}$ denote the set of translations and transliteration choices corresponding to a given source word s_j where l is the number of translations found in dictionary for s_j . The set of possible translations for the entire query Q is $T = \{tr(s_i), tr(s_j), tr(s_k)\}$. As explained earlier, out of all possible combinations of translations, the most probable translation of query is the combination which has the maximum number of co-occurrences in the corpus. However, this approach is not only computationally expensive but may also run into data sparsity problem. Hence, we use a page-rank style iterative disambiguation algorithm proposed by Christof Monz *et. al.* [2] which examines pairs of terms to gather partial evidence for the likelihood of a translation in a given context.

3.1 Iterative Disambiguation Algorithm

Given a query Q and the translation set T , a co-occurrence graph is constructed as follows: the translation candidates of different query terms are linked together. But, no edges exist between different translation candidates of the same query

term as shown in Figure 3 (a). In the above graph, $w^n(t|s_i)$ is the weight associated with node t at iteration n and denotes the probability of the candidate t being the right translation choice for the input query word s_i . A weight $l(t, t')$, is also assigned to each edge (t, t') which denotes the strength of relatedness between the words t and t' .

Initially, all the translation candidates are assumed to be equally likely.

Initialization step:

$$w^0(t|s_i) = \frac{1}{|tr(s_i)|} \quad (1)$$

After initialization, each node weight is iteratively updated using the weights of nodes linked to it and the weight of link connecting them.

Iteration step:

$$w^n(t|s_i) = w^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} l(t, t') * w^{n-1}(t'|s) \quad (2)$$

where s is the corresponding source word for translation candidate t' and $inlink(t)$ is the set of translation candidates that are linked to t . After each node weight is updated, the weights are normalized to ensure they all sum to one.

Normalization step:

$$w^n(t|s_i) = \frac{w^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w^n(t_m|s_i)} \quad (3)$$

Steps 2 and 3 are repeated iteratively till they converge approximately. Finally, the two most probable translations for each source word are chosen as candidate translations.

Link-Weights Computation The link weight, which is meant to capture the association strength between the two words (vertices), could be measured using various functions. In this work, we use two such functions: *Dice Coefficient (DC)* and *Point-wise Mutual Information (PMI)*.

PMI [3] is defined as follows:

$$l(t, t') = PMI(t, t') = \log_2 \frac{p(t, t')}{p(t) * p(t')} \quad (4)$$

where $p(t, t')$ is the joint probability of t and t' i.e. the probability of finding the terms t and t' together, in a given context, in the corpus. $p(t)$ and $p(t')$ are the marginal probabilities of t and t' respectively i.e. the probability of finding these terms in the entire corpus.

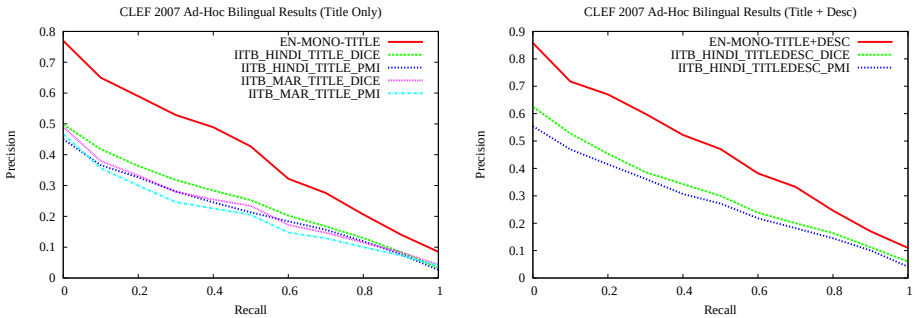
DC is defined as follows:

$$l(t, t') = DC(t, t') = \frac{2 * freq(t, t')}{freq(t) + freq(t')} \quad (5)$$

where $freq(t, t')$, $freq(t)$ and $freq(t')$ are the combined and individual frequency of occurrence of terms t and t' respectively. For computing $freq(t, t')$, which is needed for both the measures, we consider co-occurrences at the document level.

Table 2. CLEF 2007 Ad-Hoc Monolingual and Bilingual Overall Results (Percentage of monolingual performance given in brackets below the actual numbers)

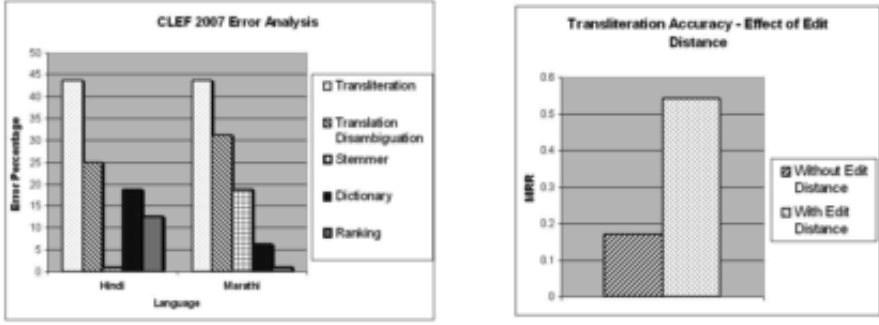
Run Desc.	Title Only					
	MAP	R-Precision	P@5	P@10	P@20	Recall
EN-MONO-TITLE	0.3856	0.3820	0.5440	0.4560	0.3910	81.40%
IITB_HINDI_TITLE_DICE	0.2366 (61.36%)	0.2468 (64.60%)	0.3120 (57.35%)	0.2920 (64.03%)	0.2700 (69.05%)	72.58% (89.16%)
IITB_HINDI_TITLE_PMI	0.2089 (54.17%)	0.2229 (58.35%)	0.2800 (51.47%)	0.2640 (57.89%)	0.2390 (61.12%)	68.53% (84.19%)
IITB_MAR_TITLE_DICE	0.2163 (56.09%)	0.2371 (62.07%)	0.3200 (58.82%)	0.2960 (64.91%)	0.2510 (64.19%)	62.44% (76.70%)
IITB_MAR_TITLE_PMI	0.1935 (50.18%)	0.2121 (55.52%)	0.3240 (59.56%)	0.2680 (58.77%)	0.2280 (58.31%)	54.07% (66.42%)
Run Desc.	Title + Description					
	MAP	R-Precision	P@5	P@10	P@20	Recall
EN-MONO-TITLE+DESC	0.4402	0.4330	0.5960	0.5040	0.4270	87.67%
IITB_HINDI_TITLEDESC_DICE	0.2952 (67.06%)	0.3081 (71.15%)	0.3880 (65.10%)	0.3560 (70.63%)	0.3150 (73.77%)	76.55% (87.32%)
IITB_HINDI_TITLEDESC_PMI	0.2645 (60.08%)	0.2719 (62.79%)	0.3760 (63.09%)	0.3500 (69.44%)	0.2950 (69.09%)	72.76% (82.99%)

**Fig. 3.** CLEF 2007 Ad-Hoc Monolingual and Bilingual Precision-Recall Curves

4 Experiments and Results

We used *Trec Terrier* [4] as the monolingual English IR engine and Okapi BM25 as the ranking algorithm. The details of the topics and document set are given in [6]. The documents were indexed after stemming (using Porter Stemmer) and stop-word removal. We used the Hindi and Marathi stemmers and morphological analyzers developed at CFILT, IIT Bombay for stemming the topic words. For each of the Title and Title + Description runs, we tried DC and PMI for calculating the link weight. This gave rise to four runs for Hindi. For Marathi, due to resource constraints, we could not carry out the Title + Description run and only did the Title run.

We use the following standard measures [5] for evaluation: Mean Average Precision (MAP), R-Precision, Precision at 5, 10 and 20 documents and Recall. We also report the percentage of monolingual English retrieval achieved for each performance figure. The overall results are tabulated in Table 2 and the corresponding precision-recall curves appear in Figure 3.



(a) Percentage of errors module-wise

(b) Effect of edit distance

Fig. 4. CLEF 2007 Analysis of Results

4.1 Discussion

In agreement with the results reported by Christof Monz *et. al.* [2], we observe that, as an association measure, DC consistently performs better than PMI. One reason for this behavior is that DC, when compared to PMI which uses a logarithmic function, is more sensitive to slight variations in frequency counts. Figure 3 (b) depicts this phenomenon where we vary the joint frequency count $f(t_i, t_j)$, keeping the individual term frequencies $f(t_i), f(t_j)$ constant.

The output of the transliteration module is a list of transliterations ranked by edit distance. We evaluated its accuracy on the CLEF 2007 topic words which had to be actually transliterated. We used the standard *Mean Reciprocal Rank (MRR)* metric for evaluation which is defined as: $MRR = \sum_{i=1}^N \frac{1}{Rank(i)}$ where $Rank(i)$ is the rank of the correct transliteration in the ranked list. We observe that the simple rule based transliteration works quite well with an MRR of 0.543 *i.e.* on an average it outputs the correct translation at rank 2. The addition of edit distance module drastically improves the accuracy as shown in Fig. 4 (b).

4.2 Error Analysis

We performed an error analysis of all the queries. We categorized these errors based on the modules in which the errors occurred. A graph depicting the percentage error contributions by various modules for each language is shown in Figure 4 (a).

For both Hindi and Marathi, the largest error contribution is due to *Devanagari to English Transliteration*. Since, we only use a single grapheme mapping, it is difficult to capture different spelling variations of a Devanagari word in English. For instance, while transliterating the word “**क़ीन**” (Queen), the correct mapping for the letter ‘क’ is ‘qa’. However, since we only have a single mapping, ‘क’ is mapped to ‘ka’ and hence it doesn’t get rightly transliterated into *Queen*. The next major source of error is the *Translation Disambiguation* module. Since we have considered document-level co-occurrence, many unrelated words also

usually co-occur with the given word due to which the DC/PMI score increases. Other important sources of error were language specific resources like Stemmer and Bi-lingual dictionaries.

5 Conclusion

We presented our Hindi to English and Marathi to English CLIR systems developed for the CLEF 2007 Ad-Hoc Bilingual Task. Our approach is based on query translation using bi-lingual dictionaries. Transliteration of words which are not found in the dictionary is done using a simple rule based approach. It makes use of the corpus to return the 'k' closest possible English transliterations of a given Hindi/Marathi word. Disambiguating the various translations/transliterations is performed using an iterative page-rank style algorithm which is based on term-term co-occurrence statistics.

Based on the current experience, we plan to explore the following directions in future: In transliteration, instead of a single rule for each letter, multiple rules could be considered. Calculating the joint frequency count at a more finer level like sentence or n-gram window instead of document-level. To improve ranking, the terms in the final translated query could be augmented with weights.

References

1. Gusfield, D.: Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge (1997)
2. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language information retrieval. In: SIGIR 2005, pp. 520–527. ACM Press, New York (2005)
3. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley-Interscience, New York (1991)
4. 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, pp. 517–519. Springer, Heidelberg (2005)
5. Yates, R.B., Neto, B.R.: Modern Information Retrieval. Pearson Education, London (2005)
6. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)

Amharic-English Information Retrieval with Pseudo Relevance Feedback

Atelach Alemu Argaw

Department of Computer and System Sciences, Stockholm University/KTH
`atelach@dsv.su.se`

Abstract. We describe cross language retrieval experiments using Amharic queries and English language document collection. Two monolingual and eight bilingual runs were submitted with variations in terms of usage of long and short queries, presence of pseudo relevance feedback (PRF), and approaches for word sense disambiguation (WSD). We used an Amharic-English machine readable dictionary (MRD), and an online Amharic-English dictionary for lookup translation of query terms. Out of dictionary Amharic query terms were considered as possible named entities, and further filtering was attained through restricted fuzzy matching based on edit distance which is calculated against automatically extracted English proper names. The obtained results indicate that longer queries tend to perform similar to short ones, PRF improves performance considerably, and that queries tend to fare better with WSD rather than using maximal expansion of terms by taking all the translations given in the MRD.

1 Introduction

Amharic is a Semitic language that is spoken in Ethiopia by an approximated 20-30 million people. It is a syllabic language, and uses a unique script which originated from the Ge'ez alphabet. 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.

Amharic has a complex but fairly structured morphological properties. To give some highlights: Amharic has a rich verb morphology which is based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. A significantly large part of the vocabulary consists of verbs, which exhibit different morphosyntactic properties based on the arrangement of the consonant-vowel patterns. Amharic nouns can be inflected for gender, number, definiteness, and case, although gender is usually neutral. Adjectives behave in the same way as nouns, taking similar inflections, while prepositions are mostly bound morphemes prefixed to nouns.

Eight bilingual (Amharic-English) and two monolingual (English) experiments are reported in this paper. One of the monolingual English experiments used short queries containing the title and description fields of the topic sets, while the other used long queries that contained title, description, and narrative

fields of the topics. Two of the eight bilingual retrieval experiments conducted used short Amharic queries while the remaining six used long ones. The experiments also differed from one another in terms of the word sense disambiguation (WSD) method used and the use of pseudo relevance feedback (PRF) in order to expand query terms. For indexing and retrieval, the Lemur toolkit for language modeling and information retrieval¹ was used.

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 different steps taken in the query analysis. Section 3 describes how out of dictionary terms were handled, followed by approaches for word sense disambiguation in section 4. Section 5 discusses pseudo relevance feedback, and section 6 presents details about the designed experiments and the obtained results. These results are discussed and future directives are given in the last section.

2 Query Analysis

The query analysis involves transliteration, stemming, stop word removal, look up translation, and fuzzy matching. Each of these processes are described in more detail in this section.

2.1 Transliteration

The Amharic queries were written in the Amharic script *fidel*. For ease of use and compatibility purposes, the text was transliterated to an ASCII representation using SERA². The transliteration was done using a file conversion utility called g2³ which is available in the LibEth⁴ package.

2.2 Stemming

We used an in-house developed software for stemming the Amharic query terms. The stemmer is designed to reduce morphological variants of words to their citation forms as found in the machine readable dictionary (MRD). It finds all possible segmentations of a given word according to 65 inflectional morphological rules of the language. Derivational variants are not handled since they tend to have separate entries in dictionaries. The most likely segmentation for the words is then selected based on occurrence statistics in a list of citation forms compiled from three dictionaries (Amharic-English, Amharic-Amharic, Amharic-French) and a 3.1 million words Amharic news corpus. The process is to strip off allowed

¹ <http://www.lemurproject.org/>

² SERA stands for System for Ethiopic Representation in ASCII,
<http://www.abysiniacybergateway.net/fidel/sera-faq.html>

³ g2 was made available to us through Daniel Yacob of the Ge'ez Frontier Foundation (<http://www.ethiopic.org/>)

⁴ LibEth is a library for Ethiopic text processing written in ANSI C
<http://libeth.sourceforge.net/>

prefixes and suffixes and look up the remaining stem (or alternatively, some morphologically motivated variants of it) in the list of citation forms to verify that it is a possible segmentation. Stem length is taken into consideration when further disambiguation is needed. In the cases where stems cannot be verified using the dictionary lists, frequency of occurrence in the news corpus is used to decide which segmentation to pick. See [2] for a detailed information about the stemming process.

Bigrams are handled in the same manner, but the segmentation works in such a way that prefixes are removed from the first word and suffixes from the second one only. Compound words in Amharic are usually written as two words, but there is no inflection present as the suffix of the first word and prefix of the second word in the bigram.

2.3 Lookup Translation

The query translation was done through term-lookup in an Amharic-English MRD [1] and an online dictionary⁵. The MRD contains 15,000 Amharic words and their corresponding English translations while the online dictionary contains about 18,000 entries. The lookup is done in such a way that the MRD translations are given precedence over the online dictionary translations. In using both resources, bigrams were given precedence over unigrams, and when a match is found, all senses and synonyms of the term translations as given in the dictionaries were taken.

2.4 Stop Word Removal

Non content bearing words (stop words) were removed both before and after the lookup translation. First, all bigrams were extracted and looked up. The stop words were removed after excluding the bigrams for which matches were found in the dictionaries. This was done to ensure that we are not missing any possible bigrams due to removed stop words that are part of a meaningful unit. Before translation, Amharic stop words were removed based on global and local occurrence statistics. Each word's occurrence frequency was collected from the 3.1 million words news text, and words with frequencies above 5,000 were considered to be stop words and are removed from the terms list. The remaining words were further checked by looking at their occurrence frequency in the 50 queries used. If they occur more than 15 times, they were also removed. The later stop word removal handled non content bearing words that are present in queries such as 'find', 'document', 'relevant' etc, which tend to have low occurrence frequencies in the news corpus.

English stop words were removed after the lookup translation. We used an English stop words list that comes with the Lemur toolkit, which is also used during the indexing of the English document collection.

⁵ <http://www.amharicdictionary.com/>

3 Fuzzy Matching for Out of Dictionary Terms

Amharic query terms that are most likely to be named entities were selected automatically for fuzzy matching. Such words are query words that are not removed as stop words but for which no bigram or unigram match is found in both dictionaries. The unsegmented word form was retained for fuzzy matching and very commonly occurring noun prefixes and suffixes are stripped off. Prefixes such as 'be', 'ye', 'ke', and 'le', were removed when they are attached preceding a word and suffixes 'oc', 'oc-n', 'oc-na', 'oc-n-na' when they appear as the word endings.

Automatically extracting named entities for Amharic is difficult compared to that of English since proper names in Amharic scripts are not capitalized. Hence, we implemented a very simple and straight forward proper name extraction utility for English. The extracted English proper names were then used for the subsequent process of fuzzy matching. An edit distance based fuzzy matching was done for the Amharic out of dictionary query terms that were selected to be possible named entities. Restricting the fuzzy matching to the extracted English proper names is believed to increase precision of the matches, while it lowers recall. We further restricted the fuzzy matching to contain terms with very high similarity levels only by setting the maximum allowed edit distance to be 2. Amharic terms for which no fuzzy match is found were removed while the shortest edit distance or preferred match is taken to be the English equivalent proper name for those words for which matches are found through the fuzzy matching. The preferred match is the match for which a predefined character in the Amharic word as given by the transliteration system SERA corresponds to a specific one in English. For example the Amharic transliteration 'marc' would have a 0 edit distance with the English proper name 'Marc' since the fuzzy matching is case insensitive. But the English word 'March' which has an edit distance of 1 with the Amharic word 'marc' would be preferred since the Amharic 'c' in SERA corresponds to the sound 'ch' in English.

4 Word Sense Disambiguation

During the lookup translation using both dictionaries, all the senses given in the dictionaries for each term's translation were taken. In such a case, where there is no sense disambiguation and every term is taken as a keyword, we consider the queries to be 'maximally expanded' with all available senses. The sense disambiguation in this case is left to be implicitly handled by the retrieval process. Some of the experiments discussed in the section below used the 'maximally expanded' set of translated keywords. Another set of experiments made use of only the first translation given in the dictionaries. Such an approach is an attempt to a very simplified and 'blind' word sense disambiguation, with the assumption that the most common sense of a word tends to be the first one on the list of possible translations given in dictionaries. A manual sense disambiguation was also done for comparison, to determine the effect of optimal WSD in the case of MRD based cross language information retrieval (CLIR). Two of the reported experiments made use of the manually disambiguated set of keywords.

5 Pseudo Relevance Feedback

Pseudo Relevance Feedback (PRF) is a method of automatic local analysis where retrieval performance is expected to improve through query expansion by adding terms from top ranking documents. An initial retrieval is conducted returning a set of documents. The top n retrieved documents from this set are then assumed to be the most relevant documents, and the query is reformulated by expanding it using words that are found to be of importance (high weights) in these documents. PRF has shown improved IR performance, but it should also be noted that there is a risk of query drift in applying PRF[4]. Four of the experiments used PRF by including the 20 highest weight terms from the top ranking 20 documents, with a positive coefficient⁶ of 0.5.

6 Experiments and Results

For indexing and retrieval, the Lemur toolkit for language modeling and information retrieval was used. The selection of this tool was primarily to try out language modeling approaches in Amharic-English cross language IR. We found that it was difficult to find optimal settings for the required smoothing parameters in the time frame allocated for this project, hence we reverted to the vector space models. Stop words were removed, and the Porter stemmer was used for stemming during indexing. Both features are available through the toolkit.

In information retrieval overall performance is affected by a number of factors, implicitly and explicitly. To try and determine the effect of all factors and tune parameters universally is a very complicated task. In attempting to design a reasonably well tuned retrieval system for Amharic queries and English document collections, our efforts lie in optimizing available resources, using language specific heuristics, and performing univariate sensitivity tests aimed at optimizing a specific single parameter while keeping the others fixed at reasonable values. In these experiments, we tried to see the effects of short queries vs. long queries, the use of PRF, and the effect of taking the first translation given versus maximally expanding query terms with all translations given in dictionaries.

What we refer to as long queries consisted of the title, description, and narrative fields of the topics, while short queries consisted of title and description fields. In the long queries, we filtered out the irrelevant information from the narrative fields, using cue words for Amharic. Amharic has the property that the last word in any sentence is always a verb, and Amharic verbs have negation markers as bound morphemes that attach themselves as prefixes onto the verbs. This property of Amharic has helped us in automatically determining whether or not a sentence in the narrative field of the topics is relevant to the query. Some of the sentences in the narrative fields of the topics describe what should not be included or is not relevant for the query at hand. If we include all the sentences

⁶ The coefficient for positive terms in (positive) Rocchio feedback.

in the narrative fields, such information could possibly hurt performance rather than boost it. Therefore we looked at the last word in each Amharic sentence in the narrative field and removed those that have ending verbs marked for negation. Examples of such words used include 'ayfelegum', 'aydelum', 'aynoracewm' representing negations of words like 'needed', 'necessary', etc.

6.1 Designed Experiments

Eight bilingual experiments were designed. Run 1 is the experiment where we used maximally expanded long queries while Run 2 supplemented these queries with PRF. In Run 3, maximally expanded short queries were used while Run 4 supplemented them with PRF. Run 5 used long queries with word sense disambiguation using the first-translation-given approach, and Run 6 supplemented them with PRF. Run 7 also used long queries but with manual word sense disambiguation and Run 8 supplemented them with PRF.

6.2 Results

The results obtained for the experiments discussed above are given in tables 1 and 2. Table 1 summarizes the results for the eight bilingual runs by presenting 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 number of relevant documents for each query) documents retrieved (R-Precision). Table 2 gives a summary similar to that of Table 1 for the monolingual English runs that were performed for comparison purposes.

Table 1. Summary of results for the bilingual runs

	<i>Relevant-tot</i>	<i>Relevant-retrieved</i>	<i>Avg Precision</i>	<i>R-Precision</i>
Run 1	2247	880	7.77	8.78
Run 2	2247	951	10.5	10.44
Run 3	2247	873	7.71	8.21
Run 4	2247	943	10.97	10.57
Run 5	2247	868	8.29	10.17
Run 6	2247	1030	11.75	12.87
Run 7	2247	1002	9.75	10.85
Run 8	2247	1104	12.92	13.3

Table 2. Summary of results for the monolingual English runs

	<i>Relevant-tot</i>	<i>Relevant-retrieved</i>	<i>Avg Precision</i>	<i>R-Precision</i>
Run 0	2247	1399	22.84	24.47
Run L	2247	1435	24.05	25.49

7 Discussion and Future Directives

Stemming plays a crucial role in MRD based CLIR since whether we would find the correct match in the dictionary depends on how well the stemmer does. The Amharic topic set contains 990 unique words from a total words count of 1892. We found direct matches in the MRD for only 120 unique words out of the 990 leaving 770 for further processing by the stemmer. The stemmer correctly segments 462 of the 770 unique words giving an accuracy of 60%. 144 terms (19%) were wrongly segmented while the remaining 164 terms (21%) were left unsegmented. Here it should be noted that some of the 164 unsegmented words are actually citation forms of words that are not found in the dictionaries.

Incorrectly segmented terms introduce noise through matches in the dictionaries that are irrelevant to the query. To take an example, the term 'bebali' which should be segmented as 'be-(bali)' is wrongly segmented as 'be-(bal)-i'. Correct segmentation would have allowed the term 'bali' to be passed to the fuzzy matching module which would have correctly matched it to the English named entity 'Bali'. The incorrectly segmented stem 'bal' finds a match in the dictionary and is translated as 'husband'. Such wrong segmentations lead to the inclusion of keywords that are irrelevant to the query. This kind of cases emphasize the need to optimize the stemmer's performance.

The limited number of entries in the dictionaries also contributes to decreased performance due to loss of content bearing words. 183 of the unique Amharic query terms that have no match in the dictionaries were assumed to be named entities. Of these, only a fraction are named entities, the rest are unmatched Amharic words. The amount of entries in the two dictionaries utilized is 15,000 and 18,000 with possible overlaps, and the chances of finding no match is high. In order to cater for the fact that we have too many candidate named entities, the fuzzy matching is restricted to English proper names only, a very high similarity requirement was set, and is supplemented by language specific heuristics. We intend to investigate approaches to bootstrap a named entity recognizer for Amharic, especially following the approaches discussed for Arabic by [5], as well as using a more sophisticated named entity recognizer for English to extract as many named entities as possible, rather than restrict it to proper names only.

As can be seen in the results presented above, the best retrieval performance obtained was from the manually disambiguated word senses, followed by the first-translation-given approach, while the maximal expansion comes last. Long queries, that are believed to carry more information since they have a lot more keywords, were expected to perform much better than the shorter queries, but the results show that they have comparable performance. The automatic filtering of sentences in the narrative fields for long queries performed very well, removing all non-relevant sentences. Although that is the case, most of the additional information gained by using the long queries was a repetition to what is already been available in the short ones, except for a few additions. Using the narrative field also boosts negative impact through wrong segmentation and lookup. In depth analysis of a larger set of queries might shade some light into the positive and negative impact.

The use of PRF in all cases showed a substantial increase in performance. Given that the original retrieval precision is very low, it is very encouraging to see that PRF helps in boosting performance even in such cases. We plan to further pursue using PRF, and tuning parameters pertaining to PRF.

The fact that manual WSD gave the best results and that blindly picking the first translation given has better performance than maximal MRD expansion of query terms motivates us to put more effort in investigating approaches to automatic WSD. Given the resource limitations, the best approach is most likely to use target language document collection and contextual collocation measures for sense disambiguation. We intend to investigate further approaches presented in [3] as well as experiment with a few more collocation measures.

Although the results obtained are indicative of the facts presented above, the experiments are too limited to draw any conclusions. Large scale experiments using a larger set of queries and data set including those from previous years of CLEF ad hoc tasks will be designed in order to give the results more statistical significance. The relatively low precision levels are also issues we plan to investigate further by taking a closer look at the indexing and retrieval experiments.

References

1. Aklilu, A.: Amharic English Dictionary. Mega Publishing Enterprise, Ethiopia (1981)
2. Argaw, A.A., Asker, L.: An amharic stemmer: Reducing words to their citation forms. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic, June 2007, pp. 104–110. Association for Computational Linguistics (2007)
3. Argaw, A.A., Asker, L., Cöster, R., Karlgren, J., Sahlgren, M.: Dictionary-based amharic-french information retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Muller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 83–92. Springer, Heidelberg (2006)
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
5. Shaalan, K., Raza, H.: Person name entity recognition for arabic. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic, June 2007, pp. 17–24. Association for Computational Linguistics (2007)

Indonesian-English Transitive Translation for Cross-Language Information Retrieval

Mirna Adriani, Herika Hayurani, and Syandra Sari

Faculty of Computer Science

University of Indonesia

Depok 16424, Indonesia

mirna@cs.ui.ac.id, {heha51,sysa51}@ui.edu

Abstract. This is a report on our evaluation of using some language resources for the Indonesian-English bilingual task of the 2007 Cross-Language Evaluation Forum (CLEF). We chose to translate an Indonesian query set into English using machine translation, transitive translation, and parallel corpus-based techniques. We also made an attempt to improve the retrieval effectiveness using a query expansion technique. The result shows that the best retrieval performance was achieved by combining the machine translation technique and the query expansion technique.

1 Introduction

To participate in the bilingual 2007 Cross Language Evaluation Forum (CLEF) task, i.e., the Indonesian-English CLIR, we needed to use language resources to translate Indonesian queries into English. However, there were not many language resources that were available on the Internet for free. We sought out for some language resources that can be used for the translation process. 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. This lead us to exploring other possible approaches such as using machine translation techniques [3], parallel corpus-based techniques, and also transitive translation techniques. Previous work has demonstrated that parallel corpus could be used as a way to find word pairs in different languages [4, 5, 6]. The word pairs could then be used to translate the queries from one language to be used to retrieve documents in another language. If such resource is not available, another possibility is by translating through some other language, known as pivot language, that has more language resources [3, 7, 8].

2 The Query Translation Process

As a first step, we manually translated the original CLEF query set from English into Indonesian. We then translated the resulting Indonesian queries back into English using machine translation technique, transitive queries technique, and the parallel corpus. For the machine translation technique, we translate the Indonesian queries into English using the available machine translation on the Internet. The transitive

technique uses German and French as the pivot languages. So, Indonesian queries are translated into French and German using bilingual dictionaries, then the German and French queries are translated into English using other dictionaries. The third technique uses a parallel corpus to translate the Indonesian queries. We created a parallel corpus by translating all the English documents in the CLEF collection into Indonesian using a commercial machine translation software called *Transtool*¹. We then created the English queries by taking a certain number of terms from certain number of documents that appear in the top document list.

2.1 Query Expansion Technique

Adding the translated queries with relevant terms (known as query expansion) has been shown to improve CLIR effectiveness [1, 3]. One of the query expansion techniques is called the *pseudo relevance feedback* [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 applied this technique in this work. To choose the relevant terms from the top ranked documents, we used the $tf*idf$ term weighting formula [9]. 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 technique, transitive technique, and the parallel corpus (Figure 1). The machine translation technique translates the Indonesian queries into English using *Toggletext*², a machine translation that is available on the Internet.

The transitive technique translates the Indonesian queries into English through German and French as the pivot languages. The translation is done using a dictionary. All of the Indonesian words are translated into German or French if they are found on the bilingual dictionaries, otherwise they are left in the original language.

In our experiments we took several approaches to handling transitive translation such as using English sense words found in either German or French dictionary (Union); and using only English sense words that appear in both German and French dictionaries (Intersection).

For the parallel corpus-based technique, we used pseudo translation to get English words using Indonesian queries. First, an Indonesian query is used to retrieve the top N Indonesian documents through an IR system. Next, we identify English documents that are parallel (paired) to these top N Indonesian documents. From the top N English documents, we created the equivalent English query based on the top T terms that have highest $tf-idf$ scores [9].

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

² See <http://www.toggletext.com/>

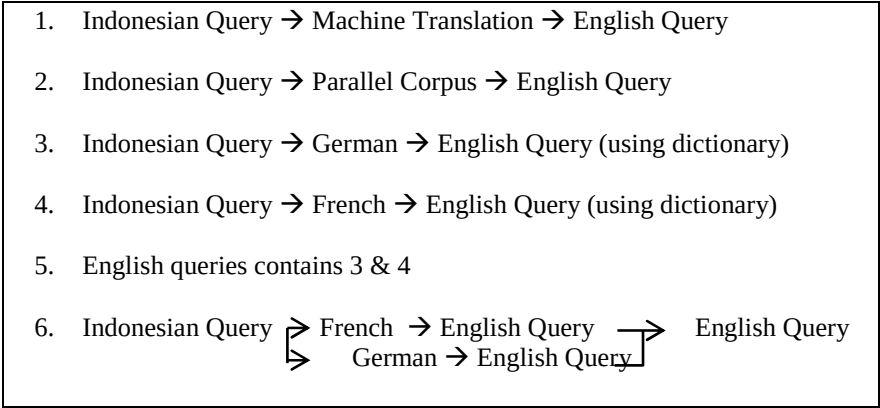


Fig. 1. The translation techniques that are used in the experiments

We then applied a pseudo relevance-feedback query-expansion technique to the queries that were translated using the three techniques above. In these experiments, we used Lemur³ information retrieval system, which is based on a language model, to index and retrieve the documents. In these experiments we also use the synonym operators to handle the translation words that are found in the dictionaries. The synonym operator gives the same weights to all the words inside it.

4 Results

Our work focused on the bilingual task using Indonesian queries to retrieve documents in the English collections. Our experiments contain official runs that have identification labels and non-official runs that do not have identification labels. Table 1-6 shows the result of our experiments.

The retrieval performance of the title-based translation queries dropped 15.59% below that of the equivalent monolingual retrieval (see Table 1). The retrieval performance of using a combination of query title and description dropped 15.72% below that of the equivalent monolingual queries.

Table 1. Mean Average Precision (MAP) of the monolingual runs of the title and combination of title and description topics and their translation queries using the machine translation

Query	Monolingual	Machine Translation (MT)
Title (depok.uiqttoggle)	0.3835	0.3237 (-15.59%)
Title + Description (depok.uiqtdtoggle)	0.4056	0.3418 (-15.72%)

³ See <http://www.lemurproject.org/>

The retrieval performance of the title-based translation queries dropped 1.64% below that of the equivalent monolingual retrieval (see Table 2) after applying the query expansion technique to the translated queries. It is increased the average precision retrieval performance by 13.95% compared to the machine translation only. However, applying query expansion to the combination of the query title and description achieves 4.38% below that of the equivalent monolingual queries. It increases the average retrieval precision of the machine translation technique by 11.34%.

Table 2. Mean Average Precision (MAP) of the monolingual runs of the title and combination of title and description topics and their translation queries using the machine translation and query expansion techniques

Query	Monolingual	MT + QE
Title (depok.uiqttogglefb10d10t)	0.3835	0.3772 (-1.64%)
Title + Description (depok.uiqtdtogglefb10d10t)	0.4056	0.3878 (-4.38%)

Table 3. Mean Average Precision (MAP) of the monolingual runs of the title and combination of title and description topics and their translation queries using transitive translation (Indonesian queries are translated to English queries via German only and via French only)

Query	Monolingual	Transitive Translation
Title + Description (via French only- depok.uiqtdfrsyn)	0.4056	0.2697 (-33.50%)
Title + Description (via German only-depok.uiqtddesyn)	0.4056	0.2878 (-29.04%)
Title + Description (via German and French)	0.4056	0.2710 (-33.18%)

The result of using the transitive translation technique for the combination of the title and description queries is shown in Table 3. Translating the queries into English using French as the pivot language decreased the mean average precision by 33.50% compared to the monolingual queries. Translating the Indonesian queries into English using German as the pivot language decreased the mean average precision by 29.04% compared to the monolingual queries. Translating Indonesian queries into English queries using two pivot languages decreases the mean average precision by 33.18% compared to the monolingual queries.

The transitive translation technique was applied for translating the Indonesian queries into English via German and French. All the English terms that were derived from the German and French words were taken based on the union and the intersection between the two sets. Adding Indonesian words that could not be translated into English resulted in a drop of the average precision by 34.56% compared to the

Table 4. Mean Average Precision (MAP) of the monolingual runs of the title and combination of title and description topics and their translation queries using transitive translation (Indonesian queries are translated to English queries via German)

Query	Monolingual	Transitive Translation
Title + Description	0.4056	0.2878 (-29.04%)
Title + Description + QE (depok.uiqtddesynfb10d10t)	0.4056	0.3342 (-17.60%)
Title + Description + QE(depok.uiqtddesynfb10d10t)	0.4056	0.3460 (-14.69%)
Title + Description + QE(depok.uiqtddesynfb5d10t)	0.4056	0.3432 (-15.38%)

Table 5. Mean Average Precision (MAP) of the monolingual runs of the title and combination of title and description topics, their translation queries using transitive translation (Indonesian queries are translated to English queries via German and French), and applying the query expansion

Query	Monolingual	Transitive Translation
Title + Description (uiqtintersectionunionsyn)	0.4056	0.2831 (-30.20%)
Title + Description + QE (depok. uiqtdintersectionunion- synf b5d10t)	0.4056	0.3437 (-15.26%)
Title + Description + QE (depok.uiqtdintersectionunionsynf b10d10t)	0.4056	0.3297 (-18.71%)
Title + Description (Union)	0.4056	0.2710 (-33.18%)
Title + Description (Intersection & add untranslated Ind terms)	0.4056	0.2654 (-34.56%)

equivalent monolingual queries. Applying the query expansion technique (see Table 5) to the resulting English queries resulted in retrieval performance that is 15-33% below the equivalent monolingual queries. The best result of using query expansion for the translated queries was obtained by taking the intersection approach, which resulted in retrieval performance 15.26% lower than that of the equivalent monolingual queries.

When the query expansion technique was applied to the translated queries resulted from using German as the pivot language the average retrieval performance dropped by 14-17% compared to the equivalent monolingual queries (see Table 4).

Table 6. Mean Average Precision (MAP) of the monolingual runs of the title and combination of title and description topics and their translation queries using parallel corpus and query expansion

Query	Monolingual	Parallel Corpus
Title + Description	0.4056	0.0374 (-90.77%)
Title + Description + QE (5 terms from 5 terms)	0.4056	0.0462 (-88.60%)

Next, we obtained the English translation of the Indonesian queries using the parallel corpus-based technique. The pseudo translation that we applied to the Indonesian queries was done by taking the English documents that are parallel with the Indonesian documents marked as relevant to the Indonesian queries by the information retrieval system. We then took the top T English terms as the English queries that had the highest weights within the top N documents. The result (see Table 6) shows that the mean average precision dropped by 90.77% of the equivalent monolingual queries. The query expansion technique that was applied to the English queries only increased the mean average precision by 2.17%. The result of the parallel corpus-based technique was very poor because the Indonesian version of the English documents in the corpus was of poor quality, in terms of the accuracy of the translation.

The retrieval performance of the transitive translation using one language, i.e. German, is better than using two languages, i.e., German and French. Translating Indonesian queries through German resulted in fewer definitions or senses than through French, meaning that the ambiguity of translating through Indonesian-German-English is less than that of translating through Indonesian-French-English.

5 Summary

Our results demonstrate that the retrieval performance of queries that were translated using a machine translation technique for Bahasa Indonesia achieved the best retrieval performance compared to the transitive technique and the parallel corpus technique. However, two of the machine translation techniques for Indonesian and English produced different results. Even though the best result was achieved by translating Indonesian queries into English using one machine translation technique; another machine translation technique that was used for creating parallel corpus produced poor results. The result of using the transitive translation technique showed that by using only one pivot language, the retrieval performance of the translated queries was better than using two pivot languages.

The query expansion that is applied to the translated queries improves the retrieval performance of the translated queries. Even though the transitive technique performance was not as good as the machine translation technique, it can be considered as a viable alternative method for the translation process, especially for languages that do not have many available language resources such as Bahasa Indonesia.

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: CLEF 2000 Working Note Workshop, Portugal (2000)
3. Ballesteros, L.A.: Cross Language Retrieval via transitive translation. In: Croft, W.B. (ed.) *Advances in Information Retrieval: Recent Research from the CIIR*, pp. 203–234. Kluwer Academic Publishers, Dordrecht (2000)
4. Chen, J., Nie, J.: Automatic Construction of Parallel English-Chinese Corpus for Cross-Language Information Retrieval. In: *Proceedings of the 6th Conference on Applied Natural Language Processing*, pp. 90–95. ACM Press, New York (2000)
5. Larenko, V., Choquette, M., Croft, W.B.: Cross-Lingual Relevance Models. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 175–182. ACM Press, New York (2002)
6. Nie, J., Simard, M., Isabelle, P., Durand, R.: Cross-Language Information Retrieval Based on Parallel Text and Automatic Mining of Parallel Text from the Web. In: *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York (1999)
7. Gollins, T., Sanderson, M.: Improving Cross Language Retrieval with Triangulated Retrieval. In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 90–95. ACM Press, New York (2004)
8. Lehtokangas, R., Airio, E., Jarvelin, K.: Transitive Dictionary Translation Challenges Direct Dictionary Translation in CLIR. *Information Processing and Management: An International Journal* 40(6), 973–988 (2004)
9. Salton, G., McGill, M.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York (1983)

Robust Retrieval Experiments at the University of Hildesheim

Ben Heuwing and Thomas Mandl

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 2007. We applied a system previously tested for ad-hoc retrieval. Experiments were focused on the effect of blind relevance feedback and named entities. Experiments for mono-lingual English and French are presented. A topic analysis of post run results shows directions for future research.

1 Introduction

We intended to provide a baseline for the robust task at CLEF 2007 [3]. Our basic system was used at CLEF campaigns previously [1].

For the baseline experiments, we optimized blind relevance feedback (BRF) parameters. The underlying basic retrieval engine of the system is the open source search engine Apache Lucene.

2 System Description

Five runs for the English and three for the French monolingual data were submitted. The results for both test and training topics are shown in table 1 and 2, respectively.

Optimization of the blind feedback parameters on the English training topics of 2006 showed the best results when the query was expanded with 30 terms from the top10 documents and the query-expansion was given a relative weight of 0.05 compared to the rest of the query. The same improvements (compared to the base run) can be seen on a smaller scale for the submitted runs. Named entities were extracted using the Lingpipe Java library for linguistic analysis and were indexed in separate fields of the same index. Assigning a higher weight to these named entities during the ranking process did not have an effect on the retrieval quality. For the post-experiments we tried to find an explanation for the considerable reduction of retrieval quality compared to the training runs. Removing the high field-weight for the document content compared to the document titles, which had been an optimization for the 2006 topics, brought an improvement of about 70% (MAP 0.0832 using same settings as for the submitted Run HiMoEnBrf2 without any weighting).

For the French runs, the use of a heavy-weighted (equal to the rest of the query) query expansion with 50 terms from the best five documents came out as the best blind relevance parameters – which is consistent with the results of the training topics.

Table 1. Results for Submitted Runs

<i>Run</i>	<i>Language</i>	<i>Stemming</i>	<i>BRF (weight-docs-terms)</i>	<i>NE</i>	<i>MAP</i>	<i>R-Precision</i>	<i>Precision @10</i>
HiMoEnBase	English	snowball	-	-	0.0527	0.0769	0.1060
HiMoEnBrf1	English	snowball	1.0-10-30	-	0.0580	0.0888	0.1300
HiMoEnBrf2	English	snowball	0.05-10-30	-	0.0586	0.0858	0.1090
HiMoEnBrfNe	English	snowball	0.05-10-30	1.0	0.0588	0.0858	0.1090
HiMoEnNe	English	snowball	-	2.0	0.0527	0.0769	0.1060
HiMoFrBase	French	lucene	-	-	0.2584	0.2543	0.2970
HiMoFrBrf	French	lucene	0.5-5-25	-	0.2634	0.2687	0.3080
HiMoFrBrf2	French	lucene	1.0-5-50	-	0.2731	0.2752	0.3190

Table 2. Results for Training Topics

<i>Run</i>	<i>Language</i>	<i>Stemming</i>	<i>BRF (weight-docs-terms)</i>	<i>NE</i>	<i>MAP</i>
HiMoEnBase	English	snowball	-	-	0.1634
HiMoEnBrf1	English	snowball	1.0-10-30	-	0.1489
HiMoEnBrf2	English	snowball	0.05-10-30	-	0.1801
HiMoEnBrfNe	English	snowball	0.05-10-30	1.0	0.1801
HiMoEnNe	English	snowball	-	2.0	0.1634
HiMoFrBase	French	lucene	-	-	0.2081
HiMoFrBrf	French	lucene	0.5-5-25	-	0.2173
HiMoFrBrf2	French	lucene	1.0-5-50	-	0.2351

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 leads to a sharp drop in the performance according to this measure.

3 Topic Analysis

In a topic analysis for French, we analyzed topics for which our system failed. We defined failure with Average Precision below 1% and adopted a failure category system as suggested by Savoy [4].

Taking a closer look at these topics, the following issues can be identified:

- *Stemming* seemed to conflate several terms to a base form with potentially many meanings (e.g. “Internet” -> “intern” which is also a French word meaning internal and which occurs in many compound words, “vents” -> “ven” which can also be the stem of “vend” meaning “sell”).
- In other cases, *phrases* consisting of several words seemed to cause problems (“fuite des cerveaux”, “diseurs de bonne aventure”). The words within these phrases occur frequently in other meanings in the collection.
- The scope of some of the failed topics was very *unspecific*, while the narrative asked for more specific aspects. In addition, poor performance of several topics is due to explicit references to events of the year 1995 (the targeted news collections being from 1994).

The problems caused by stemming could perhaps be reduced by the use of stemming rules which take into account the loss of specificity for each term. Preserving phrases might also help to improve the robustness of retrieval for difficult topics. These features should be implemented and evaluated regarding their benefit for robust retrieval.

The post-run without field-weights for the English sub-task brought small improvements for most of the topics. Results improved considerably for the following failed topics: 294 308 317 325 335 338 and 341. Some of the initial failures may be due to a figurative language use (hurricane, eclipse). Some of the topics still have a low performance. We observed that BRF added many terms unrelated to the meaning of the topics.

4 Outlook

For future experiments, we intend to exploit the knowledge on the impact of named entities on the retrieval process [2] as well as selective relevance feedback strategies in order to improve robustness.

References

1. Mandl, T., Hackl, R., Womser-Hacker, C.: Robust Ad-hoc Retrieval Experiments with French and English at the University of Hildesheim. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 127–128. Springer, Heidelberg (2007)
2. 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), Santa Fe, New Mexico, USA, March 13.-17, pp. 1059–1064 (2005)
3. Di Nunzio, G.M., et al.: Overview of the ad-hoc Track at CLEF. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
4. Savoy, J.: Why do successful search systems fail for some topics. In: Proceedings of the 2007 ACM Symposium on Applied Computing, SAC 2007, Seoul, Korea, March 11 - 15, 2007, pp. 872–877. ACM Press, New York (2007)

SINAI at CLEF Ad-Hoc Robust Track 2007: Applying Google Search Engine for Robust Cross-Lingual Retrieval

F. Martínez-Santiago, A. Montejo-Ráez, and M.A. García-Cumbreras

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

Abstract. We report our web-based query generation experiments for English and French collections in the Robust task of the CLEF Ad-Hoc track. We continued with the approach adopted in the previous year, although the model has been modified. Last year we used Google to expand the original query. This year we create a new expanded query in addition to the original one. Thus, we retrieve two lists of relevant documents, one for each query (the original and the expanded one). In order to integrate the two lists of documents, we apply a logistic regression merging solution. The results obtained are discouraging but the failure analysis shows that very difficult queries are improved by using both queries instead of the original query. The problem is to decide when a query is very difficult.

1 Introduction

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[1]. Due to the multilinguality of the web, we have assumed that this could be extended to additional languages, though the smaller amount of web non-English pages could be a major obstacle. Therefore, we have used Google in order to expand the query in a similar way [2], but instead of replacing the original query by the expanded query, we have executed both queries (the original and expanded one). For each query we have obtained a list of relevant documents. Thus, we need to combine the retrieval results from these two independent list of documents. This is a similar problem to the so called collection fusion problem [3], but we have not several collections: there is only one collection but two lists of relevant documents. The question is how should we calculate of the score of each document in the final resulting list. Given a query, in order to integrate the information available about the relevance of every retrieved document, we have applied a model based on logistic regression. Logistic regression has been used successfully in multilingual scenarios [4],[5].

2 Query Expansion with Google Search Engine

This section describes the process for generating a new query using expansion by the Google search engine. To this end, we have selected a random sample

document. The following fields correspond to the document with identification number 10.2452/252-ah from the English collection.

```
<title>pension schemes in europe </title>

<desc>find documents that give information about current pension
systems and retirement benefits in any european country. </desc>

<narr>relevant documents will contain information on current pension
schemes and benefits in single european states. information of
interest includes minimum and maximum ages for retirement and the way
in which the retirement income is calculated. plans for future
pension reform are not relevant. </narr>
```

These fields have been concatenated into one single text and all contained nouns, noun phrases and prepositional phrases have been extracted by means of *TreeTagger*. TreeTagger is a tool for annotating text with part-of-speech and lemma information which has been developed at the Institute for Computational Linguistics of the University of Stuttgart¹.

Once nouns and phrases are identified they are taken to compose the query, preserving phrases thanks to Google's query syntax.

```
documents 'pension schemes' benefits retirement information
```

The former string is passed to Google and the snippets (small fragment of text from the associated web page result) of the top 100 results are joined into one single text from which, again, phrases are extracted with their frequencies to generate a final expanded query. The 20 most frequent nouns, noun phrases and prepositional phrases from this generated text are replicated according to their frequencies in the snippets-based text and then normalized to the minimal frequency in those 20 items (i.e. normalized according to the least frequent phrase among the top ones). The resulting query is shown below:

```
pension pension pension pension pension pension pension pension
pension pension pension pension pension pension pension pension
pension pension pension benefits benefits benefits benefits benefits
benefits benefits benefits benefits retirement retirement retirement
retirement retirement retirement retirement retirement retirement
retirement retirement retirement age age pensions occupational
occupational occupational occupational schemes schemes schemes
schemes schemes schemes schemes schemes schemes schemes schemes
schemes schemes schemes schemes schemes regulations information
information information information information scheme scheme
disclosure disclosure pension schemes pension schemes pension
schemes pension schemes pension schemes pension schemes pension
schemes pension schemes pension schemes pension schemes pension
schemes pension schemes retirement benefits schemes members members
occupational pension schemes occupational pension schemes
occupational pension schemes retirement benefits retirement benefits
disclosure of information
```

French documents have been processed in a similar way, but using the OR operator to join found phrases for the generated Google query. This has been done due to the smaller number of indexed web pages in French language. Since we expect to recover 100 snippets, we have found that with this operator this is

¹ Available at <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

possible, despite low quality texts been included to produce the final expanded query.

The next step is to execute both original and Google queries on the Lemur information retrieval system. The collection dataset has been indexed using Lemur IR system². It is a toolkit 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. 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. In these experiments we have used Okapi as weighting function [6].

Finally, we have to merge both lists of relevant documents. [7],[8] propose a merging approach based on logistic regression. Logistic regression is a statistical methodology for predicting the probability of a binary outcome variable according to a set of independent explanatory variables. The probability of relevance to the corresponding document D_i will be estimated according to four parameters: the score and the ranking obtained by using the original query, and the score and the ranking by means of the Google-based query (see equation 1). Based on these estimated probabilities of relevance, the list of documents will be interleaved making up an unique final list.

$$Prob[D_i \text{ is rel} | rank_{org_i}, rsv_{org_i}, rank_{google_i}, rsv_{google_i}] = \frac{e^{\alpha + \beta_1 \cdot \ln(rank_{org_i}) + \beta_2 \cdot rsv_{org_i} + \beta_3 \cdot \ln(rank_{google_i}) + \beta_4 \cdot rsv_{google_i}}}{1 + e^{\alpha + \beta_1 \cdot \ln(rank_{org_i}) + \beta_2 \cdot rsv_{org_i} + \beta_3 \cdot \ln(rank_{google_i}) + \beta_4 \cdot rsv_{google_i}}} \quad (1)$$

The coefficients α , β_1 , β_2 , β_3 and β_4 are unknown parameters of the model. When fitting the model, usual methods to estimate these parameters are maximum likelihood or iteratively re-weighted least squares methods.

In order to fit the underlying model, training set (topics and their relevance assessments) must be available for each monolingual collection. Since there are relevance assessments for English and French, we have made the experiments for these languages only. For Portuguese we have reported only the base case (we have not used Google queries for such language).

3 Results

As Tables 1 and 2 show, the results are disappointing. For training data, Google queries improve both the m.a.p. and the geometric precision in both languages, English and French. But this good behavior disappears when we apply our approach on test data. Of course, we expect that precision for test data gets worse regarding training data, but we think that the difference in precision is excessive.

In order to evaluate the impact of Google approach we have made an analysis of English results. We suspect that Google approach is only a good idea when

² <http://www.lemurproject.org/>

Table 1. Results for English data. Google approach is the result obtained by merging original queries and Google queries. Base results are those obtained by means of original queries only.

Approach	Collection	map	gm-ap
Google	training	0.29	0.12
Base	training	0.26	0.10
Google	test	0.34	0.12
Base	test	0.38	0.14

Table 2. Results for French data. Google approach is the result obtained by merging original queries and Google queries. Base results are those obtained by means of original queries only.

Approach	Collection	map	gm-ap
Google	training	0.28	0.10
Base	training	0.26	0.12
Google	test	0.30	0.11
Base	test	0.31	0.13

queries are very hard. Thus, we have made two clusters of queries. The first cluster is formed by “easy” queries: the mean average precision is higher than 10 percentage points. The second cluster is formed by the “difficult queries”. Since the MAP is different for original and Google results we have defined the clusters by using the MAP of original queries, test data (we have made the same analysis taking into account the MAP of Google test and training data in order to create the clusters, and the results are virtually the same that we report here. We don’t report all the cases because of the length of the paper).

Table 3. Results for English data when the MAP is higher than 10 percentage points (69 queries)

Approach	Collection	map	gm-ap
Google	training	0.35	0.19
Base	training	0.36	0.10
Google	test	0.38	0.14
Base	test	0.43	0.17

The most interesting result reported in Table 3 and Table 4 is related to the geometric precision. Google approach overcomes base approach in 2-3 points if the MAP is lower than 10 percent of the points. Thus, we think that the Google approach is a useful strategy only for very difficult queries. Now the question is, how do we know that a query is difficult? This is a key issue on robust IR systems. We found that this strategy used on all queries usually makes the system get worse average results, so a study on how to apply the method in a different way should be undertaken.

Table 4. Results for English data when the MAP is lower than 10 percentage points (31 queries)

Approach	Collection	map	gm-ap
Google	training	0.08	0.10
Base	training	0.05	0.08
Google	test	0.07	0.10
Base	test	0.07	0.08

4 Conclusions and Future Work

We have reported our experimentation for the Ad-Hoc Robust Multilingual track CLEF task involving web-based query generation for English and French collections. The generation of a final list of results by merging search results obtained from two different queries has been studied. These two queries are the original one and the one generated from Google results. Both lists are joined by means of logistic regression, instead of using an expanded query as we did last year. The results are disappointing. While results for training data are very promising, there is no improvement for test data. Nevertheless, the analysis of the results shows that there is a improvement if only difficult queries are considered. The main problem is that difficult queries are identified by evaluating the system on relevance assessments.

Acknowledgments

This work has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id.514 granted by the University of Jaén.

References

1. Kwok, K.L., Grunfeld, L., Lewis, D.D.: TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In: Proceedings of TREC'3, vol. 500-215, pp. 247–255. NIST Special Publication (1995)
2. Martínez-Santiago, F., Montejó-Ráez, A., García-Cumbreras, M.A., Ureña-López, L.A.: SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: Query Expansion using the Google Search Engine Evaluation of Multilingual and Multi-modal Information Retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
3. Voorhees, E., Gupta, N.K., Johnson-Laird, B.: The Collection Fusion Problem. In: Proceedings of the 3th Text Retrieval Conference TREC-3, vol. 500-225, pp. 95–104. NIST Special Publication (1995)
4. Martínez Santiago, F., Ureña López, L.A., Martín-Valdivia, M.T.: A merging strategy proposal: The 2-step retrieval status value method. Information Retrieval 9, 71–93 (2006)

5. Savoy, J.: Combining Multiple Strategies for Effective Cross-Language Retrieval. *Information Retrieval* 7, 121–148 (2004)
6. Robertson, S.E., Walker, S.: Okapi-Keenbow at TREC-8. In: *Proceedings of the 8th Text Retrieval Conference TREC-8*, vol. 500-246, pp. 151–162. NIST Special Publication (1999)
7. Calvé, A., Savoy, J.: Database merging strategy based on logistic regression. *Information Processing & Management* 36, 341–359 (2000)
8. Savoy, J.: Cross-Language information retrieval: experiments based on CLEF 2000 corpora. *Information Processing & Management* 39, 75–115 (2003)

Improving Robustness Using Query Expansion

Angel F. Zazo, José L. Alonso Berrocal, and Carlos G. Figuerola

REINA Research Group – University of Salamanca
C/ Francisco Vitoria 6-16, 37008 Salamanca, Spain
{zazo,berrocal,figue}@usal.es

Abstract. This paper describes our work at CLEF 2007 Robust Task. We have applied local query expansion using windows of terms, but considering different measures of robustness during the training phase in order to optimize the performance: MAP, GMAP, MMR, GS@10, P@10, number of failed topics, number of topics below 0.1 MAP, and number of topics with P@10=0. The results were not disappointing, but no settings were found that simultaneously improved all measures. A key issue for us was to decide which set of measures we had to select for optimization.

This year all our runs also gave good rankings, both base runs and expanded ones. However, our expansion technique does not improve significantly the retrieval performance. At TREC and CLEF Robust Tasks other expansion techniques have been used to improve robustness, but results were not uniform. In conclusion, regarding robustness the objective must be to make good information retrieval systems, rather than to tune some query expansion techniques.

1 Introduction

This year our research group has participated in two tracks at CLEF 2007: 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 offered at TREC 2003, 2004 and 2005 for monolingual retrieval [1, 2, 3], and at CLEF 2006 for monolingual, bilingual and multilingual retrieval [4]. This year only monolingual (English, French and Portuguese) and bilingual (English to French) subtasks were offered. Our research group has participated in all the subtasks. For a complete description of this task, please, see the CLEF 2007 Ad-hoc Track Overview, also published in this volume.

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. However, 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, so at CLEF 2007 other criteria of robustness have been

suggested: MAP, GMAP, P@10, number of failed topics, number of topics below 0.1 MAP, and number of topics with P@10=0. In our experiments we have also considered two other user-related measures: the Generalized Success@10 (GS@10) [5], and the mean reciprocal rank (MRR). Both indicate the rank of the top retrieved relevant document.

At TREC, query expansion using both the document collection and external corpora (Web and other document collections) was the best approach for robust retrieval. At CLEF 2006 Robust Task our research group obtained very good results applying local query expansion using windows of terms [6]. This year we have used the same expansion technique, but taking the new measures into account.

Our main focus was monolingual retrieval. The steps followed are explained below. For bilingual retrieval experiments we used machine translation (MT) programs to translate topics into document language, and then a monolingual retrieval was implemented.

2 Experiments

At CLEF 2006 our monolingual runs gave very good rankings, therefore we decided to use this year the same information retrieval system and also the same settings for our experiments. We used the well-known vector space model, using the **dnu-ntc** term weighting scheme. For documents, letter *u* stands for the pivoted document normalization [7]: we adjusted *pivot* to the average document length and *slope* set to 0.1 for all the collections. We decided to remove the most frequent terms in each collection, those which had a document frequency of at least a quarter of the number of documents in the collection.

Last year we verified that stemming not only does not deteriorate the retrieval performance of hard topics for all the collections, but there was also a significant improvement for the French and Italian collections. So this year we also decided to apply stemming for all the collections, including the Portuguese one. For English we used the Porter stemmer, and for French and Portuguese the stemmers from the University of Neuchatel at <http://www.unine.ch/info/clef/>.

It should be noted that we automatically removed certain phrases from the descriptions and narratives of the topics, such as “Find documents that ...”, “Les documents pertinents relatent ...” or “Encontrar documentos sobre ...”.

The last step was to apply local query expansion using windows of terms [6]. This technique uses co-occurrence relations in windows of terms from the first retrieved documents to build a thesaurus to expand the original query. Taking into account the new criteria of robustness, a lot of tests were carried out to obtain the best performance using the training collections. We used long (title + description) and short (title) queries in the tests. The results were not disappointing, but no settings were found that simultaneously improved all measures. A key issue for us was to decide which set of measures we had to select for optimization. At CLEF 2006 and 2007 Workshops there were discussions about this problem. Finally, we decided to select the settings that improve the greatest number of measures.

For English the highest improvement achieved with this expansion technique was obtained by using a distance value of 1, taking the first 15 retrieved documents to build the thesauri, and adding about 10 terms to the original query. For French, the highest improvement was achieved by using a distance value of 1, taking the first 20 retrieved documents, and adding 40 terms to the original query.

For Portuguese we decided to use the best combination obtained last year for the Spanish experiments, due two reasons. First, the Portuguese language is more similar to Spanish than to English or to French. Second, the average number of terms per sentence in the Portuguese collection is very similar to the Spanish one. We use a distance value of 2, taking the first 10 documents, and adding 30 terms to the original query.

Table 1. Results of the runs submitted at CLEF 2007 Robust Task

		Basis	Expansion*	Basis	Expansion*	Basis
		t	t	td	td	tdn
English	MAP	0.3226	0.3205	0.3897	0.3855	0.3897
	GMAP	0.1190	0.1045	0.1850	0.1762	0.1850
	(*)Settings	MRR	0.5602	0.5379	0.6922	0.6792
	for expansion:	GS@10	0.7613	0.7219	0.8506	0.8422
	distance=1	P@10	0.3200	0.3240	0.3620	0.3640
	docs=15	# failed	5	5	5	5
	terms=10	# <0.1 MAP	16	20	7	8
		# P@10=0	16	23	10	11
						10
French	MAP	0.3382	0.3481	0.3773	0.3804	0.3773
	GMAP	0.0940	0.0947	0.1289	0.1218	0.1289
	(*)Settings	MRR	0.5749	0.5972	0.6564	0.6564
	for expansion:	GS@10	0.7555	0.7445	0.7940	0.7959
	distance=1	P@10	0.3710	0.3740	0.4140	0.4280
	docs=20	# failed	9	9	8	9
	terms=40	# <0.1 MAP	18	19	12	12
		# P@10=0	23	24	19	18
						19
Portuguese	MAP	0.3387	0.3533	0.4083	0.4121	0.4140
	GMAP	0.0825	0.0911	0.1369	0.1301	0.1287
	(*)Settings	MRR	0.5711	0.5950	0.6286	0.6273
	for expansion:	GS@10	0.7307	0.7277	0.7855	0.7718
	distance=2	P@10	0.3013	0.3027	0.3320	0.3347
	docs=10	# failed	15	12	10	10
	terms=30	# <0.1 MAP	28	29	22	26
		# P@10=0	36	39	29	30
						30
EN → FR	MAP	0.3035	0.3278	0.3385	0.3455	0.3583
	GMAP	0.0821	0.0872	0.1005	0.0997	0.1228
	(*)Settings	MRR	0.5819	0.6084	0.6219	0.6164
	for expansion:	GS@10	0.7555	0.7580	0.7833	0.7769
	distance=1	P@10	0.3242	0.3535	0.3770	0.3870
	docs=20	# failed	9	9	9	9
	terms=40	# <0.1 MAP	16	16	15	14
		# P@10=0	22	20	19	18
						16

For the bilingual experiments, the CLIR system was the same as that used in monolingual retrieval. A previous step was carried out before searching, to translate English topics into French. We used three MT programs: L&H Power Translator Pro 7.0, Systran¹ and Reverso². For each topic we combined the terms of the translations in a single topic: this is another expansion process, although in most cases the three translations were identical. Finally, a monolingual retrieval was performed. The local query expansion using co-occurrence based thesauri built with terms windows was also applied.

3 Results

As regards test and training topics, five runs were submitted for each subtask and topic language. Table 1 shows the results of the test runs. We can see that local query expansion using windows of terms does not improve performance for all measures. However, our expansion technique does not deteriorate significantly any of them; therefore we consider that it is a good expansion technique. This year all our runs also gave good rankings, both base runs and expanded ones, taking into account that MAP and GMAP were the measures of robustness used in the task.

4 Conclusions

At CLEF 2006 Robust Task our research group obtained very good results applying local query expansion using windows of terms for monolingual retrieval. This year at CLEF 2007 all our runs also gave very good rankings. MAP and GMAP measurements were again used this year. In our tests we applied different measures of robustness during the training phase in order to optimize the performance. A key issue for us was to decide which set of measures we had to select for optimization. At TREC conferences and at CLEF 2006 and 2007 Workshops there were discussions about this problem. Tomlinsom [5] introduced alternative ideas: “measures based on the first relevant item reflect robustness”. He deals with *primary recall measures*, based on the retrieval of the first relevant document for a topic (GS10 and MRR), and *secondary recall measures*, based on the retrieval of the additional relevant documents for a topic after the first one (MAP, GMAP and P@10).

At TREC and CLEF Robust Tasks, query expansion using both the document collections and/or external corpora (Web and other document collections) was the most used approach for robust retrieval, using MAP and GMAP measurements for robustness. In our experiments using query expansion at CLEF 2006 and 2007 we have verified that it is easy to find some settings to improve the secondary recall measures, but that improving primary ones is a difficult task. We also have verified that some query expansion techniques improved MAP and GMAP, but hard topics behaved differently. Since other authors had obtained

¹ <http://www.systransoft.com>

² <http://www.reverso.net>

the same results [1, 8], we wonder whether the robustness of a system should be measured using primary recall measures.

On the other hand, there are alternative approaches for robustness. Some researchers have used morphological techniques such as stemming, with very good results for some languages [9, 10]. This year we used a simple document retrieval system, and we looked for a good document-query weighting scheme as the basis for our next expansion experiments. We checked whether stopword removal or stemming processes improved the system robustness. We saw that stemming, in general, improves the performance of hard topics. With these settings our base runs gave good rankings. In conclusion, regarding robustness the objective must be to make good information retrieval systems, rather than to tune some query expansion techniques.

For the bilingual English to French subtask, collecting terms from some translations of a topic seems to be a good technique. Our mandatory run was the best run in the subtask.

References

1. Voorhees, E.M.: Overview of the TREC 2003 robust retrieval track. In: The Twelfth Text REtrieval Conference (TREC 2003), vol. 500-255, pp. 69–77. NIST Special Publication (2003)
2. Voorhees, E.M.: Overview of the TREC 2004 robust retrieval track. In: The Thirteenth Text REtrieval Conference (TREC 2004), Gaithersburg, Maryland, November 16–19, vol. 500-261, pp. 70–79. NIST Special Publication (2004)
3. Voorhees, E.M.: Overview of the TREC 2005 robust retrieval track. In: The Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, Maryland, November 15–18. NIST Special Publication 500-266 (2005)
4. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad hoc track overview. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 21–34. Springer, Heidelberg (2007)
5. Tomlinson, S.: Comparing the robustness of expansion techniques and retrieval measures. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 129–136. Springer, Heidelberg (2007)
6. Zazo, A.F., Alonso Berrocal, J.L., Figuerola, C.G.: Local query expansion using terms windows for robust retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 145–152. Springer, Heidelberg (2007)
7. Singhal, A., Buckley, C., Mitra, M.: Pivoted document length normalization. In: Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21–29 (1996)
8. Tomlinson, S.: Early precision measures: implications from the downside of blind feedback. In: SIGIR, pp. 705–706 (2006)
9. Tomlinson, S.: Sampling precision to depth 10000: evaluation experiments at CLEF 2007. In: CLEF 2007 Workshop Working Notes (2007)
10. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007: Ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 13–32. Springer, Heidelberg (2008)

English-to-French CLIR: A Knowledge-Light Approach through Character N -Grams Alignment

Jesús Vilares¹, Michael P. Oakes², and Manuel Vilares³

¹ Dept. of Computer Science, University of A 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@sunderland.ac.uk`

³ Dept. of Computer Science, University of Vigo
Campus As Lagoas s/n, 32004 - Ourense (Spain)
`vilares@uvigo.es`

Abstract. This paper describes an extension of our work presented in the robust English-to-French bilingual task of the CLEF 2007 workshop, a knowledge-light approach for query translation in Cross-Language Information Retrieval systems. Our work is based on the direct translation of character n -grams, avoiding the need for word normalization during indexing or translation, and also dealing with out-of-vocabulary words. Moreover, since such a solution does not rely on language-specific processing, it can be used with languages of very different nature even when linguistic information and resources are scarce or unavailable. The results obtained have been very positive, and support the findings from our previous English-to-Spanish experiments.

1 Introduction

This work is an extension of our proposal originally developed for the robust English-to-French bilingual task of the CLEF 2007 workshop [1]. It consists of a knowledge-light approach for query translation in Cross-Language Information Retrieval (CLIR) systems based on the direct translation of character n -grams. This proposal itself can be considered as an extension of the previous work by [2].

The use of overlapping character n -grams both as indexing and translation units provides a means to normalize word forms. In addition, the approach supports the handling of out-of-vocabulary words and the management of languages of very different nature without further processing. Moreover, such a 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.

Since the architecture of our system has been described in depth in a previous CLEF publication [3], this paper focuses on the work performed after the workshop. The paper is structured as follows: firstly, Sect. 2 briefly introduces our

approach; next, Sect. 3 presents the new experiments; finally, Sect. 4 contains our conclusions and proposals for future work.

2 Description of the System

Taking as our model the system designed by JHU/APL [2], we have developed our own n -gram based retrieval system, trying to preserve the advantages of the original system but avoiding its main drawbacks.

The main difference with our proposal is the n -gram alignment algorithm, the basis of the system, which consists of two phases. In the first phase, the slowest one, the input parallel corpus is aligned at the word-level using the statistical tool GIZA++ [4], obtaining as output the translation probabilities between the different source and target language words. In our case, taking advantage of our previous experiments with English-to-Spanish [5,6], we have opted for a bidirectional alignment [7] which considers, for example, a (w_{EN}, w_{FR}) English-to-French word alignment only if there also exists a corresponding (w_{FR}, w_{EN}) French-to-English alignment. This way, subsequent processing is focused only on those words whose translation seems less ambiguous, considerably reducing the number of input word pairs to be processed —actually about 70%— and, consequently, the noise introduced in the system. This reduction allows us to greatly reduce both computing and storage resources.

Next, prior to the second phase, we have also removed those least-probable word alignments from the input (those with a word translation probability less than a threshold W , with $W=0.15$) [5,6]. Such pruning leads to a considerable reduction of processing time and storage space: a reduction of about 95% in the number of input word pairs processed.

Finally, in the second phase, n -gram translation scores are computed using statistical association measures [8], taking as input the translation probabilities previously calculated by GIZA++, and weighting the likelihood of a cooccurrence according to the probability of its containing word alignments [5,6].

For this purpose, our system employs three of the most extensively used standard measures: the *Dice coefficient* (*Dice*), *mutual information* (*MI*), and *log-likelihood* (*logl*), which are defined by the following expressions [8]:

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

$$logl(g_s, g_t) = 2 \sum_{i,j} O_{ij} \log \frac{NO_{ij}}{R_i C_j} . \quad (3)$$

3 Evaluation

In the past CLEF 2007 workshop, our group took part in the robust English-to-French bilingual task. The *robust task* is essentially an ad-hoc task which re-uses the topics and collections from past CLEF editions [9].

Unfortunately, our system could not be accurately tuned for the workshop. So, we had to use the parameters employed in our previous English-to-Spanish experiments [5,6]. Moreover, only one of the two selection algorithms available was used, the so-called *top-rank-based* algorithm. This section presents the work developed after the CLEF 2007 workshop for tuning the system for the new target language. This new experiments also include the *threshold-based* selection algorithm.

With respect to the indexing process, documents were simply split into n -grams and indexed. We used 4-grams as a compromise n -gram size [5,6]. Before that, the text was lowercased and punctuation marks were removed [2], but not diacritics. The open-source TERRIER platform [10] was used as retrieval engine with a InL2¹ ranking model [11]. No stopword removal or query expansion were applied at this point.

For querying, the source language topic² is firstly split into n -grams. Next, these n -grams are replaced by their candidate translations according to a selection algorithm, and the resulting translated topics are then submitted to the retrieval system. Two alternative selection algorithms were implemented: a *top-rank-based* algorithm, that takes the N highest ranked n -gram alignments according to their association measure, and a *threshold-based* algorithm, that takes those alignments whose association measure is greater than or equal to a threshold T .

The work presented in this paper was developed in two phases. Firstly, the *training topics* subset was used for tuning the system for the different association measures implemented: the Dice coefficient, mutual information and log-likelihood. Next, the performance was tested using the *test topics* subset.³

3.1 Tuning Runs Using the Dice Coefficient

The first tuning runs were made for the Dice coefficient and the top-rank-based selection algorithm, that is, by taking the target n -grams from the N top n -gram-level alignments with the highest association measures. Different values were tried, with $N \in \{1, 2, 3, 5, 10, 20, 30, 40, 50, 75, 100\}$. The results obtained are shown in the left hand graph of Fig. 1⁴ —notice that mean average precision (MAP) values are also given. The best results were obtained when using a limited number of translations, those obtained with $N=1$ being the best.

The next tuning runs were made for the threshold-based selection algorithm, that is, by fixing a minimal association measure threshold T . Since the Dice coefficient takes values in the range $[0..1]$, we tried different values $T \in \{0.00, 0.001, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00\}$. The results obtained are shown in the right hand graph of Fig. 1, with the best ones at $T=0.40$. Nevertheless, these results were significantly less good than those for the top-rank-based algorithm.⁵

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

² Only *title* and *description* topic fields were used in the submitted queries.

³ All these experiments must be considered as *unofficial* experiments.

⁴ Only a subset of the results are shown in order not to crowd the figures.

⁵ Two-tailed T-tests over MAPs with $\alpha=0.05$ have been used along this work.

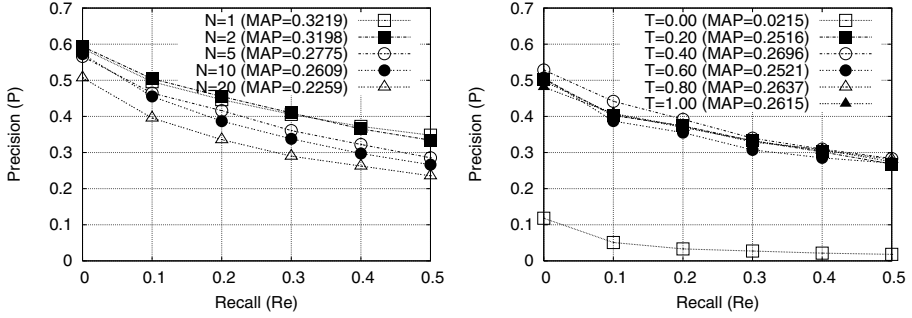


Fig. 1. Tuning precision vs. recall graphs for the Dice coefficient when using the top-rank-based (left) and threshold-based (right) selection algorithms

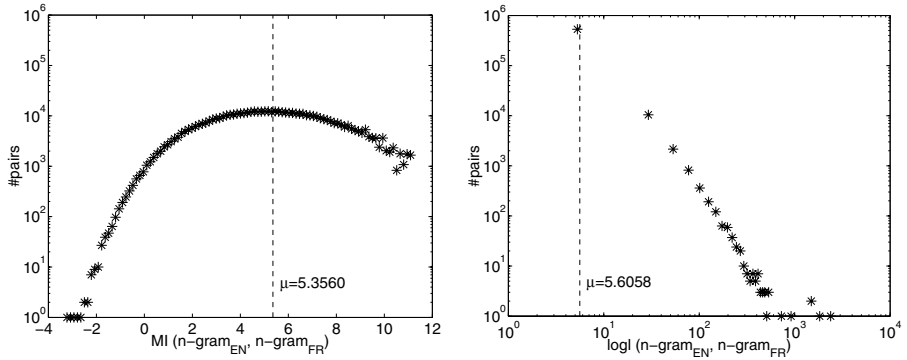


Fig. 2. Distribution of aligned n -gram pairs across their association measures: mutual information (MI, left) and log-likelihood (logl, right). Mean (μ) is also shown.

3.2 Tuning Runs Using Mutual Information

The main difference of mutual information (MI) with respect to the Dice coefficient is that the former can take any value within $(-\infty..+\infty)$ —the distribution found with our data is shown in the left hand graph of Fig. 2—, while the latter takes values within the range $[0..1]$. This had to be taken into account in order to adapt our testing methodology.

In the case of the top-rank-based selection algorithm, we continued taking the N top-ranked n -gram alignments, even if their MI value was negative. The results obtained, shown in the left hand graph of Fig. 3, and with the best performance at $N=10$, were not as good as those obtained with the Dice coefficient.

In the case of the threshold-based algorithm, we had to take into account that the range of MI values may vary considerably for each run. So, in order to homogenize the experiments, the threshold values were not fixed according to concrete values as before, but according to the following formula:

$$T_i = \mu + 0.5 \, i \, \sigma . \quad (4)$$

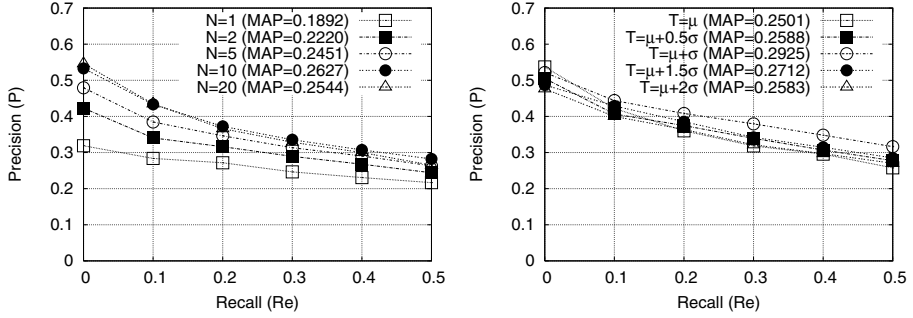


Fig. 3. Tuning precision vs. recall graphs for mutual information when using the top-rank-based (left) and threshold-based (right) selection algorithms

where T_i represents the i -th threshold —with $i \in \mathbb{N}$ —, μ represents the *mean* of the MI values of the aligned n -gram pairs, and σ represents their *standard deviation*. The resulting thresholds are as follows:

$$\mu, \mu + 0.5\sigma, \mu + \sigma, \mu + 1.5\sigma, \dots$$

The right hand graph of Fig. 3 shows the results obtained for this algorithm, which were very similar, although the best ones were obtained for $T = \mu + \sigma$. These results were better than those obtained for the top-rank-based algorithm, but this difference was not statistically significant. However, these results were still not as good as those obtained for the Dice coefficient.

3.3 Tuning Runs Using Log-Likelihood

As before, the first runs used the top-rank-based algorithm. These results, shown in the left hand graph of Fig. 4, and with the best performance at $N=2$, were similar to those obtained for the Dice coefficient.

Regarding the threshold-based selection algorithm, log-likelihood, like MI, does not have a fixed range of possible values. So, as with MI, we established the thresholds according to the *mean* and *standard deviation* of the association measures. Nevertheless, after studying the distribution of the output aligned n -gram pairs across their log-likelihood values —see right hand graph of Fig. 2—, we realized that this distribution was clearly biased towards low values just slightly less than the mean. As a consequence, we worked with varying granularities and developed the following formula for calculating the threshold values:

$$T_i = \begin{cases} \mu + 0.05 i \sigma & -\infty < i \leq 2 \\ \mu + 0.50 (i - 2) \sigma & 2 < i < +\infty \end{cases} \quad (5)$$

where, as before, T_i represents the i -th threshold —this time with $i \in \mathbb{Z}$ —, μ represents the *mean* of the log-likelihood values of the aligned n -gram pairs, and σ represents their *standard deviation*. The resulting thresholds are as follows:

$$\dots \mu - 0.05\sigma, \mu, \mu + 0.05\sigma, \mu + 0.1\sigma, \mu + 0.5\sigma, \mu + \sigma \dots$$

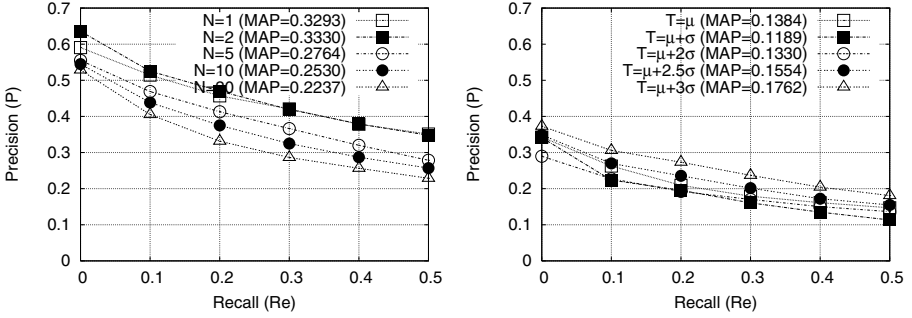


Fig. 4. Tuning precision vs. recall graphs for log-likelihood when using the top-rank-based (left) and threshold-based (right) selection algorithms

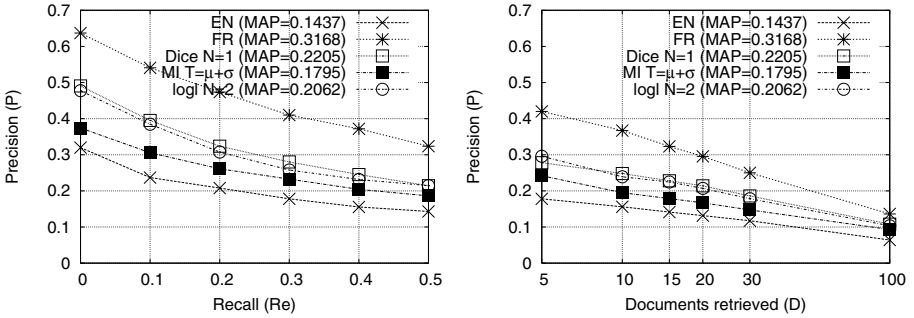


Fig. 5. Precision vs. recall graphs (left) and precision at top D documents graphs (right) for the *test topics* set

The results obtained, shown in the right hand graph of Fig. 4, and with the best performance for $T = \mu + 3\sigma$, were significantly worse than those for the top-rank-based algorithm.

3.4 Test Runs

Once the system had been tuned for the new target language, the proper tests could be performed using the *test topics* set. The best configurations found for each association measure were used in these runs:

- Dice coefficient (EN2FR Dice): top-rank-based selection algorithm ($N = 1$)
- Mutual Information (EN2FR MI): threshold-based selection algorithm ($T = \mu + \sigma$)
- Log-likelihood (EN2FR logl): top-rank-based selection algorithm ($N = 2$)

Fig. 5 presents the results obtained for the test runs with respect to two baselines: the first by querying the French index with the initial English topics split into 4-grams (EN) —allowing us to measure the impact of casual matches—, and the other obtained by querying the French index using the French topics split into 4-grams (FR) —i.e. a French monolingual run and our *ideal performance* goal.

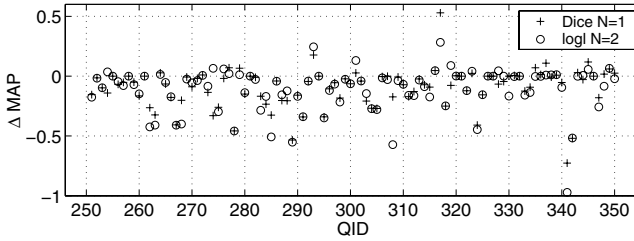


Fig. 6. MAP differences with the French monolingual run for the *test topics* set

These results showed that the Dice coefficient and the log-likelihood measure produced the best results —when using the top-rank-based algorithm. Both approaches performed significantly better than mutual information —the latter using a threshold-based algorithm. Regarding the baselines, all configurations performed significantly better than the English topics run, showing that our positive results were not coincidental. On the other hand, these results were still not as good as the French monolingual run, our *ideal performance* goal, but it must be taken into account that this approach is not still fully developed, so there is margin for improvement. Finally, Fig. 6 shows the MAP differences with the French monolingual run (FR) obtained for each topic in the case of our best configurations: the Dice coefficient (*Dice*) and the log-likelihood (*logl*) measure when using the top-rank-based algorithm.

4 Conclusions and Future Work

This work presents a knowledge-light approach for query translation in Cross-Language Information Retrieval systems based on the direct translation of character *n*-grams. The experiments shown in this paper are an extension of those performed in the robust English-to-French task of the CLEF 2007, and confirm the positive results previously obtained in our English-to-Spanish experiments [5,6], thus demonstrating the validity of our approach.

With respect to our future work, new tests with other languages of different characteristics are being prepared. We also intend to simplify the processing for reducing the computational costs even more. Finally, the employment of relevance feedback, or the use of pre or post-translation expansion techniques in the case of translingual runs [2] are also being considered.

Acknowledgments

This research has been partially funded by the European Union (FP6-045389), Ministerio de Educación y Ciencia and FEDER (TIN2004-07246-C03 and HUM2007-66607-C04), and Xunta de Galicia (PGIDIT07SIN005206PR, PGIDIT05PXIC30501PN, and *Rede Galega de Procesamento da Linguaxe e Recuperación de Información*).

References

1. <http://www.clef-campaign.org> (visited on November 2007)
2. 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)
3. Vilares, J., Oakes, M.P., Tait, J.I.: A first approach to CLIR using character n-grams alignment. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 111–118. Springer, Heidelberg (2007)
4. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1), 19–51 (2003),
<http://www.fjoch.com/GIZA++.html> (visited on November 2007)
5. Vilares, J., Oakes, M.P., Vilares, M.: A knowledge-light approach to query translation in cross-language information retrieval. In: Proc. of International Conference on Recent Advances in Natural Language Processing (RANLP 2007), pp. 624–630 (2007)
6. Vilares, J., Oakes, M.P., Vilares, M.: Character n-grams translation in cross-language information retrieval. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (eds.) NLDB 2007. LNCS, vol. 4592, pp. 217–228. Springer, Heidelberg (2007)
7. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: NAACL 2003: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54 (2003)
8. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)
9. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 ad hoc track overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
10. <http://ir.dcs.gla.ac.uk/terrier/> (visited on November 2007)
11. 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)

MIRACLE Progress in Monolingual Information Retrieval at Ad-Hoc CLEF 2007

José-Carlos González-Cristóbal^{1,3}, José Miguel Goñi-Menoyo¹,
Julio Villena-Román^{2,3}, and Sara Lana-Serrano^{1,3}

¹ Universidad Politécnica de Madrid

² Universidad Carlos III de Madrid

³ DAEDALUS - Data, Decisions and Language, S.A.

josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es

jvillena@daedalus.es, slana@diatel.upm.es

Abstract. This paper presents the 2007 MIRACLE's team approach to the Ad-Hoc Information Retrieval track. The main work carried out for this campaign has been around monolingual experiments, in the standard and in the robust tracks. The most important contributions have been the general introduction of automatic named-entities extraction and the use of wikipedia resources. For the 2007 campaign, runs were submitted for the following languages and tracks: a) Monolingual: Bulgarian, Hungarian, and Czech. b) Robust monolingual: French, English and Portuguese.

1 Introduction

The MIRACLE¹ Information Retrieval toolbox is made of basic components in a classical pipeline architecture: stemming, transformation (transliteration, elimination of diacritics and conversion to lowercase), filtering (elimination of stop and frequent words), proper nouns detection and extracting, and paragraph extracting, among others. Some of these basic components can be used in different combinations and order of application for document indexing and for query processing. Standard stemmers were used from Porter [8] for English, and from Neuchatel [11] for Hungarian, Bulgarian and Czech. In the 2007 experiments, only OR combinations of the search terms were used. The retrieval model used is the well-known Robertson's Okapi BM-25 [9] formula for the probabilistic retrieval model, without relevance feedback. Through our participation in previous campaigns, the integration procedure of the different modules is stable and, to some point, optimized. MIRACLE toolbox has already been described in previous campaigns papers [2], [3], [7].

MIRACLE makes use of its own indexing and retrieval engine, which is based on the trie data structure [1]. Tries have been successfully used by the MIRACLE team for years, as an efficient storage and retrieval of huge lexical resources, combined

¹ 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 fifth participation in CLEF.

with a continuation-based approach to morphological treatment [6]. For the 2007 campaign, runs were submitted for the following languages and tracks:

- Monolingual: Bulgarian, Hungarian, and Czech.
- Robust monolingual: French, English and Portuguese.

The most relevant work carried out in this campaign was the incorporation of modules for the recognition of named entities in the tokenizing process, besides the compiling of extended resources adequate for this task.

2 Results for the Monolingual and Robust Tasks

The following table and figure summarize the performance of our official experiments in the monolingual tasks (using the topic fields title/description).

The most relevant work carried out for the 2007 campaign was the integration of components for multilingual Named Entities Recognition. In particular, entities from Wikipedia were extracted for all languages of interest in the framework of CLEF. In

Table 1. Average precision for monolingual experiments

lang	Average Precision	Prec. at 0	Prec. At 100
BG	0.2717	0.5946	0.0531
CZ	0.3203	0.6697	0.0701
HU	0.3499	0.7672	0.987

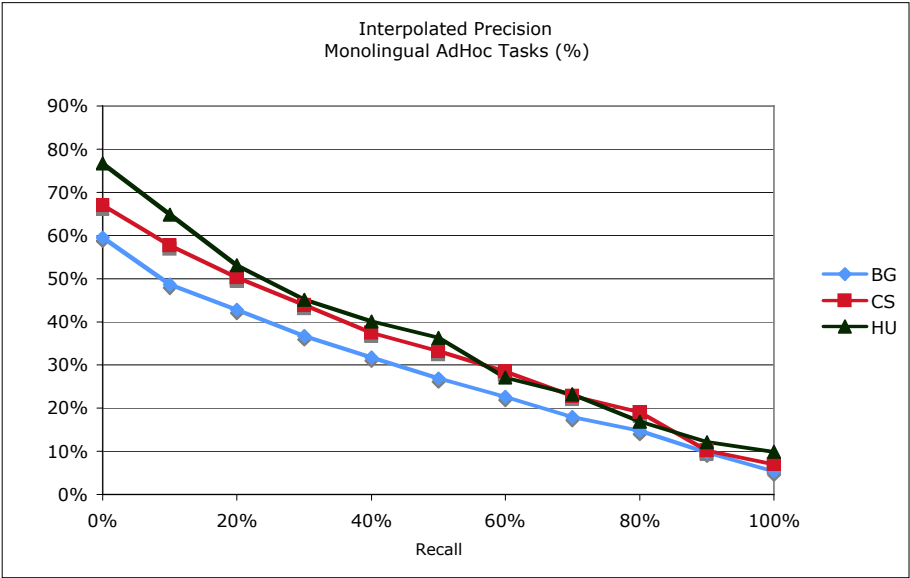


Fig. 1. Interpolated precision for monolingual experiments

the case of English, the number of entities used was above 500,000. An additional improvement was made through normalization of the recognized entities. Under this approach, the terms *United Nations*, *UN*, *U.N.* and *U. N.* were automatically substituted by an identifier associated with this international organization. For evaluation purposes, one baseline system was implemented that applied a simple Porter stemmer with lowercase reduction. The results of these experiments were fully available after the end of the campaign, and are shown in Table 2.

The results are discouraging, showing no improvement associated with the usage of these techniques, although a detailed analysis shows that the number of correctly retrieved texts is slightly higher.

Table 2. Precision figures for robust monolingual experiments in English

Run	Average Precision	Prec. at 0	Prec. at 1
Simple stemming	0.3966	0.6457	0.1688
Wikipedia-based named-entity recognition	0.3892	0.6398	0.1622
Named-entity recognition + Normalization	0.3920	0.6428	0.1658

3 Conclusions and Future Work

For the 2007 campaign, the processing scheme was maintained from previous ones, starting some improvements regarding proper nouns and entities detection and indexing. The results presented here indicate that Named Entity Recognition techniques have no impact on the TREC-based precision measures used for CLEF experiments. Although new experiments have to be conducted, it seems obvious that stemming provides a simple, fast and robust way for information retrieval of English texts. Further work includes extending this approach for languages other than English, integrated with other sets of external resources apart from Wikipedia, or through automatic learning of entities from the collections.

We still think that a high-quality entity recognition (proper nouns or acronyms for people, companies, countries, locations, and so on) can improve the precision and recall figures in some information retrieval tasks, as well as a correct recognition and normalization of dates, times, numbers, etc. In particular, such techniques can reveal a higher impact on cross-lingual tasks.

Regarding Wikipedia-based resources, three main usages are foreseen. The first one is the identification of relevant multiword expressions. The second is the expansion of acronyms. The last one is the translation of expressions between languages, to be used in future bilingual tasks. For these specific tasks, the use of multilingual thesaurus (e.g. Eurovoc) will be also generalized, as the IR platform is ready to incorporate such resources.

Acknowledgements. This work has been partially supported 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; and by the Madrid's R&D Regional Plan, by means of the MAVIR project (Enhancing the

Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

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, J.M., González, J.C., Villena, J.: MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. In: Peters, C., et al. (eds.) *Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria*. LNCS, vol. 4022, pp. 44–53. Springer, Heidelberg (2006)
3. Goñi, J.M., González, J.C., Villena, J.: Miracle's 2005 Approach to Monolingual Information Retrieval. In: *Working Notes for the CLEF 2005 Workshop, Vienna, Austria* (2005)
4. Goñi, J.M., González, J.C., Martínez, J.L., Villena, J.: MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. In: Peters, C., Clough, P., Gonzalo, J., et al. (eds.) *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004*. LNCS, vol. 3491, pp. 188–199. Springer, Heidelberg (2005)
5. Goñi, J.M., González, J.C., Martínez, J.L., Villena, J., García, A., Martínez, P., de Pablo, C., Alonso, J.: MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. In: Peters, C., Borri, F. (eds.) *Working Notes for the CLEF 2004 Workshop, Bath, United Kingdom*, pp. 141–150 (2004)
6. Goñi, J.M., González, J.C., Fombella, J.: An optimised trie index for natural language processing lexicons. *MIRACLE Technical Report*. Universidad Politécnica de Madrid (2004)
7. González, J.C., Goñi, J.M., Villena, J.: Miracle's 2005 Approach to Cross-lingual Information Retrieval. In: *Working Notes for the CLEF 2005 Workshop, Vienna, Austria* (2005)
8. Porter, M.: Snowball stemmers and resources page, <http://www.snowball.tartarus.org> [Visited 18/07/2006]
9. Robertson, S.E., et al.: Okapi at TREC-3. In: Harman, D.K. (ed.) *Overview of the Third Text REtrieval Conference (TREC-3)*, April 1995. NIST, Gaithersburg (1995)
10. Savoy, J.: Report on CLEF-2003 Multilingual Tracks. In: Peters, C., Gonzalo, J., Branschler, M., Kluck, M. (eds.) *CLEF 2003*. LNCS, vol. 3237, pp. 64–73. Springer, Heidelberg (2004)
11. University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers, etc.), <http://www.unine.ch/info/clef> [Visited 18/07/2006]

The Domain-Specific Track at CLEF 2007

Vivien Petras, Stefan Baerisch, and Maximillian Stempfhuber

GESIS Social Science Information Centre, Lennéstr. 30, 53113 Bonn, Germany
{vivien.petras, stefan.baerisch, max.stempfhuber}@gesis.org

Abstract. The domain-specific track uses test collections from the social science domain to test monolingual and cross-language retrieval in structured bibliographic databases. Special attention is given to the existence of controlled vocabularies for content description and their potential usefulness in retrieval. Test collections and topics are provided in German, English and Russian. This year, a new English test collection (from the CSA Sociological Abstracts database) was added. We present an overview of the CLEF domain-specific track including a description of the tasks, collections, topic preparation, and relevance assessments as well as contributions to the track. The track participants experimented with different retrieval models ranging from classic vector-space to probabilistic to language models. The controlled vocabularies were used for query expansion or as bilingual dictionaries for query translation.

Keywords: Information Retrieval, Evaluation, Controlled Vocabularies.

1 Introduction

The CLEF domain-specific track evaluates mono- and cross-language information retrieval on structured scientific data. A point of emphasis in this track is research on leveraging the structure of data in collections (i.e. controlled vocabularies and other metadata) to improve search. In recent years, the focus of the domain-specific data collections was on bibliographic databases in the social science domain.

The domain-specific track was established at the inception of CLEF in 2000 and was funded by the European Union from 2001-2004 [6, 7]. It is now continued at the GESIS German Social Science Information Centre (Bonn) in cooperation with the DELOS Network of Excellence on Digital Libraries.

The GIRT databases (now in version 4) are extracts from the German Social Science Information Centre's SOLIS (Social Science Literature) and SOFIS (Social Science Research Projects) databases from 1990-2000. In 2005, the Russian Social Science Corpus (RSSC) was added as a Russian-language test collection (94,581 documents), which was changed in 2006 to the INION ISISS corpus covering social sciences and economics in Russian. This year, another English-language social science collection was added. The second English collection is an extract from CSA's Sociological Abstracts providing more documents and another thesaurus to the test bed.

In addition to the four test collections, various controlled vocabularies and mappings between vocabularies were made available. As is standard for the domain-specific track, 25 topics were prepared in German and then translated into English and Russian.

2 The Domain-Specific Task

The domain-specific track includes three subtasks:

- *Monolingual retrieval* against the German GIRT collection, the English GIRT and CSA Sociological Abstracts collections, or the Russian INION ISISS collection;
- *Bilingual retrieval* from any of the source languages to any of the target languages;
- *Multilingual retrieval* from any source language to all collections / languages.

2.1 The Test Collections

In recent years, pseudo-parallel collections in German and English (GIRT) and one or two Russian test collections were provided [8, 13]. This year, only one Russian but two English collections were provided.

Every test collection is in the format of a bibliographic database (records include title, author, abstract and source information) with the addition of subject metadata from controlled vocabularies.

German

The German GIRT collection (the social science German Indexing and Retrieval Testdatabase) is now used in its forth version [5] with 151,319 documents covering the years 1990-2000 using the German version of the Thesaurus for the Social Sciences. Almost all documents contain an abstract (145,941).

English

The English GIRT collection is a pseudo-parallel corpus to the German GIRT collection, providing translated versions of the German documents. It also contains 151,319 documents using the English version of the Thesaurus for the Social Sciences, but only 17% (26,058) documents contain an abstract.

New additions this year were the documents from the social science database Sociological Abstracts from Cambridge Scientific Abstracts (CSA) with 20,000 documents, 94% of which contain an abstract. The documents were taken from the SA database covering the years 1994, 1995, and 1996. Additional to title and abstract, each document contains subject-describing keywords from the CSA Thesaurus of Sociological Indexing Terms and classification codes from the Sociological Abstracts classification.

Russian

For Russian retrieval, the INION corpus ISISS with bibliographic data from the social sciences and economics with 145,802 documents was once again used. ISISS documents contain authors, titles, abstracts (for 27% of the test collection or 39,404 documents) and keywords from the Inion Thesaurus.

2.2 Controlled Vocabularies

The GIRT collections have assigned descriptors from the GESIS IZ Thesaurus for the Social Sciences in German and English depending on the collection language. The CSA Sociological Abstracts documents contain descriptors from the CSA Thesaurus

of Sociological Indexing Terms and the Russian ISISS documents are provided with Russian INION Thesaurus terms. GIRT documents also contain classification codes from the GESIS IZ classification and CSA SA documents from the Sociological Abstracts classification. Table 1 shows the distribution of subject-describing terms per document in each collection.

Table 1. Distribution of subject-describing terms per collection

Collection	<i>GIRT-4 (German or English)</i>	<i>CSA Sociological Abstracts</i>	<i>INION ISISS</i>
Thesaurus descriptors / document	10	6.4	3.9
Classification codes / document	2	1.3	n/a

Vocabulary mappings

Additional to the “mapping table” for the German and English terms from the GESIS IZ Thesaurus for the Social Sciences, which is really a translation, mappings between the GIRT and CSA Thesauri was provided.

The vocabulary mappings used are one-directional, intellectually created term transformations between two controlled vocabularies. They can be used to switch from the subject metadata terms of one knowledge system to the other, enabling a retrieval system to treat the subject descriptions of two or more different collections as one and the same. This year’s mappings were equivalence transformations, showing only term mappings that were found to be equivalent between two different controlled vocabularies.

We provided mappings between the German Thesaurus for the Social Sciences and the English CSA Thesaurus of Sociological Indexing Terms. Since the German Thesaurus for the Social Sciences exists in an English version as well, we also provided the mapping from the English Thesaurus for the Social Sciences to the English CSA Thesaurus of Sociological Indexing Terms for monolingual retrieval.

An example for a mapping from the English Thesaurus for the Social Sciences to the English CSA Thesaurus of Sociological Indexing Terms would be:

```
<mapping>
  <original-term>agricultural area</original-term>
  <mapped-term>Rural areas</mapped-term>
</mapping>
```

This example shows that a mapping can overcome differences in technical language and the treatment of singular and plural case in different controlled vocabularies.

2.3 Topic Preparation

As is standard for the CLEF domain-specific track, 25 topics were prepared. To date, 200 topics have been created for the domain-specific track.

For topic preparation we were supported by our colleagues from the GESIS Social Science Information Centre. As a special service to the social science community in Germany, the Information Centre biannually publishes updates on new entries in the SOLIS and SOFIS databases (from which the GIRT collections were generated). The specialized updates are prepared in 28 subject categories by subject specialists working at the Centre. Topics range from general sociology, family research, women's and gender studies, international relations, research on Eastern Europe to social psychology and environmental research. An overview of the service including the 28 topics can be found at the following URL:

<http://www.gesis.org/en/information/soFid/index.htm>.

We asked our colleagues to think of between 2-5 topics related to their subject area and potentially relevant in the years 1990-2000 (the coverage of our test collections). The suggestions from 15 different colleagues were then checked according to breadth,

```
<top>
<num>192</num>
  <EN-title>System change and family planning in
  East Germany</EN-title>
  <EN-desc>Find documents describing birth trends
  and family planning since reunification in East
  Germany.</EN-desc>
  <EN-narr>Of interest are documents on demographic
  changes which have taken place after 1989 in the
  territory of the former GDR as well as the slump in
  birth numbers, decline in marriages and di-
  vorces.</EN-narr>
</top>
```

Fig. 1. Example topic in English

Table 2. Topic titles for domain-specific CLEF track 2007

Sibling relations	Class-specific leisure behaviour
Unemployed youths without vocational training	Mortality rate
German-French relations after 1945	Economic elites in Eastern Europe and Russia
Multinational corporations	System change and family planning in East Germany
Partnership and desire for children	Gender and career chances
Torture in the constitutional state	Ecological standards in emerging or developing countries
Family policy and national economy	Integration policy
Women and income level	Tourism industry in Germany
Lifestyle and environmental behaviour	Promoting health in the workplace
Unstable employment situations	Economic situations of families
Value change in Eastern Europe	European climate policy
Migration pressure	Economic support in the East
Quality of life of elderly persons	

variance from previous years and coverage in the test collections. 25 topics were selected and edited into the CLEF topic XML format. Figure 1 is an example.

All topics were created in German and then consequently translated into English and Russian.

Table 2 lists all 25 topic titles in English to give a perspective on the variance in topics.

3 Overview of the 2007 Domain-Specific Track

More details of the individual runs and methods employed can be found in the articles by the participating groups as well as their Working Notes papers.

3.1 Participants

Although 10 groups registered for the domain-specific task, only 5 groups submitted runs. Four groups have submitted papers in the track [2, 3, 10, 11]. Table 3 lists all participants.

Table 3. Domain-specific track 2007 – participants

<i>Abbreviation</i>	<i>Group Institution</i>	<i>Country</i>
Chemnitz	Media Informatics, Chemnitz Univ. of Technology	Germany
Cheshire	School of Information, UC Berkeley	USA
Xerox	Xerox Research Centre - Data Mining Group	France
Moscow	Moscow State University	Russia
Unine	Computer Science Dept., Univ. of Neuchatel	Switzerland

3.2 Submitted Runs

Experiments for all tasks (monolingual, bilingual and multilingual retrieval) were submitted to the track. Monolingual and bilingual experiments were equally often

Table 4. Submitted runs per task in the domain-specific track

Task	Runs
<i>Monolingual</i>	
- against German	13
- against English	15
- against Russian	11
<i>Bilingual</i>	
- against German	14
- against English	15
- against Russian	9
<i>Multilingual</i>	9

Table 5. Submitted runs per task and participant

Task	Participants (Runs)
<i>Monolingual</i>	
-against German	Chemnitz (3), Cheshire (2), Unine (4), Xerox (4)
- against English	Chemnitz (3), Cheshire (2), Moscow (2), Unine (4), Xerox (4)
- against Russian	Chemnitz (3), Cheshire (2), Moscow (2), Unine (4)
<i>Bilingual</i>	
- against German	Chemnitz (4), Cheshire (4), Xerox (6)
- against English	Chemnitz (3), Cheshire (4), Moscow (2), Xerox (6)
- against Russian	Chemnitz (3), Cheshire (4), Moscow (2)
<i>Multilingual</i>	Chemnitz (3), Cheshire (6)

attempted, whereas multilingual retrieval runs were only submitted by 2 groups. Russian remains slightly less popular than the other two languages. Table 4 provides the number of submitted runs per task, table 5 provides an overview over submitted runs per task per participant.

3.3 Relevance Assessments

In previous years, the domain-specific relevance assessments were administrated and overseen at least partly in-house at the Social Science Information Centre (using a self-developed Java-Swing program). This year all relevance assessments were administered and processed in the DIRECT system (Distributed Information Retrieval Evaluation Campaign Tool) provided by Giorgio M. Di Nunzio and Nicola Ferro from the Information Management Systems (IMS) Research Group at the University of Padova, Italy.

This provided tremendous assistance for the CLEF group at the Information Centre and was positively accepted by the five assessors. Some problems occurred because of bandwidth and execution problems, but overall the assessment stage went smoothly.

Documents were pooled using the top 100 ranked documents from each submission. Table 6 shows the pool sizes for each language.

Table 6. Pool sizes in the domain-specific track

German	16,288
English	17,867
Russian	14,473

For the German assessments, 652 documents per topic were judged on average and about 22% were found relevant. However, assessments vary from topic to topic. Figure 2 shows the German assessments per topic.

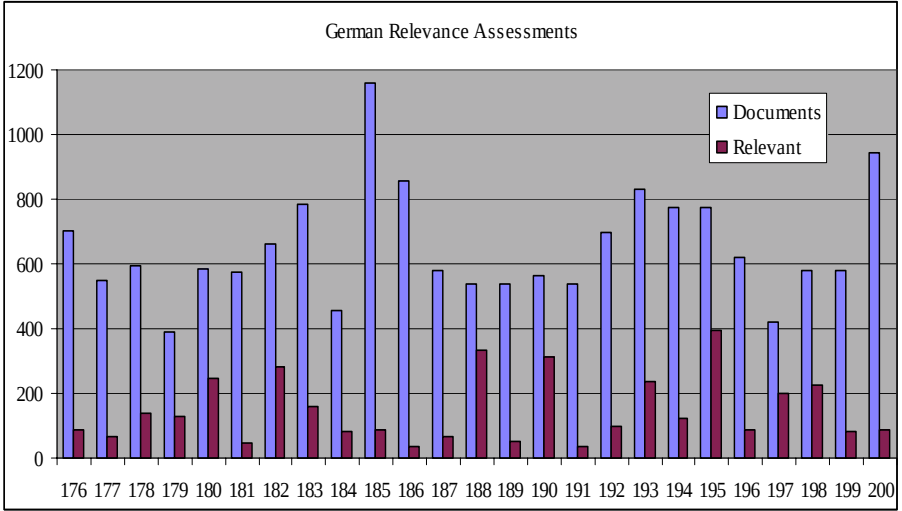


Fig. 2. German assessments per topic

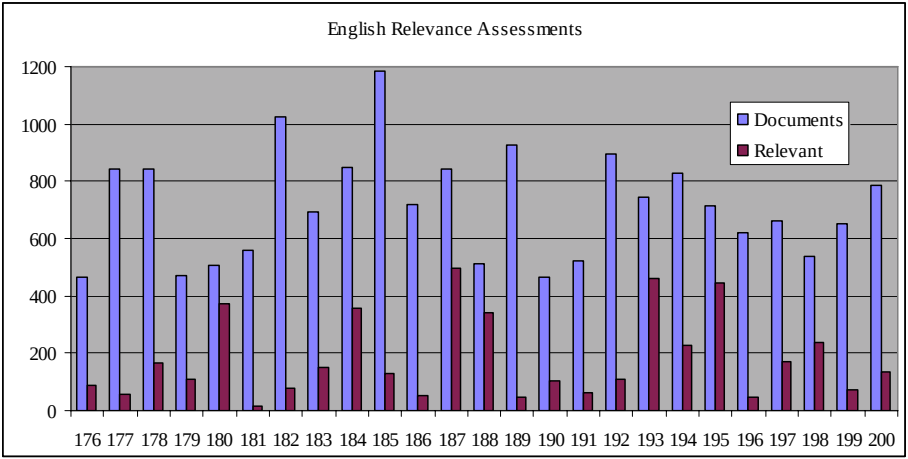


Fig. 3. English assessments per topic

For the English assessments, 715 documents per topic were judged on average and about 25% were found relevant.

For the Russian assessments, 3 topics were found to have no relevant documents in the ISISS collection: 178, 181 and 191. For the assessments, 579 documents per topic were judged and only 10% were found relevant.

Figures 3 and 4 show the English and Russian relevance assessments numbers.

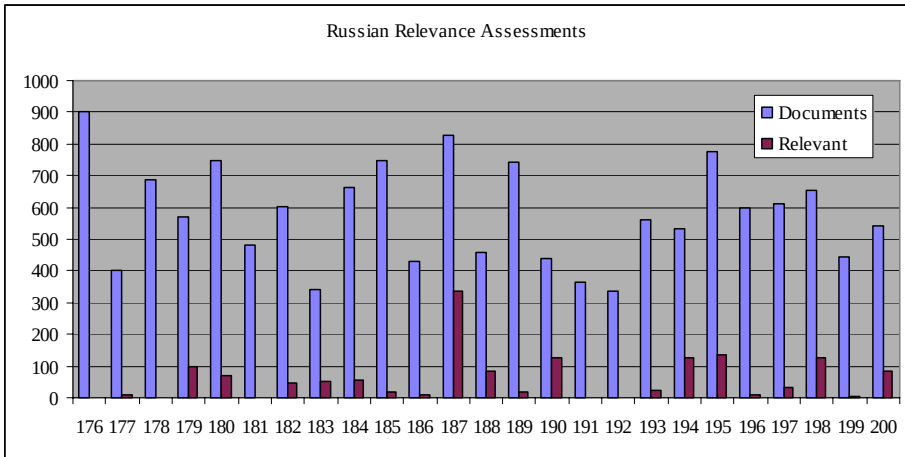


Fig. 4. Russian assessments per topic

The percentage of relevant documents per topic seems to be relatively stable across years and languages (one outlier); Table 7 provides an overview for the last 3 years.

Table 7. Percentage of relevant documents per language, 2005-2007

	English	German	Russian
Rel. docs 2007	25%	24%	10%
Rel. docs 2006	39%	26%	n/a
Rel. docs 2005	21%	20%	9%

Although the Russian collections changed from 2005 (RSSC) to 2007 (Inion), the percentage of relevant documents remains about the same. Systems retrieving against the English or German collections find more than twice as many relevant documents, with English appearing to be the “easiest” language in terms of finding relevant documents. One outlier appears in 2006, when the percentage of relevant English documents in the pool was almost 40% (in 2007: 25%). However the German percentage remained the same in 2006 and 2007.

In 2006, the German document pool contained roughly a third more documents than the English pool. With already fewer English documents available, more documents were judged relevant in this pool compared to the German documents - leading to a high percentage. In 2007, the pooled and judged relevant documents are roughly the same - with the English even having some more documents than the German pool. This might be explained by the addition of a second English collection or could simply be considered a return to a more normal distribution.

Hard and easy Topics

At first glance, several topics seem to yield particularly many relevant documents over all 3 languages despite different collections (e.g. 188, 190, 195) whereas others seem to yield particularly few (e.g. 181, 191). Figures 5 and 6 show the differences.

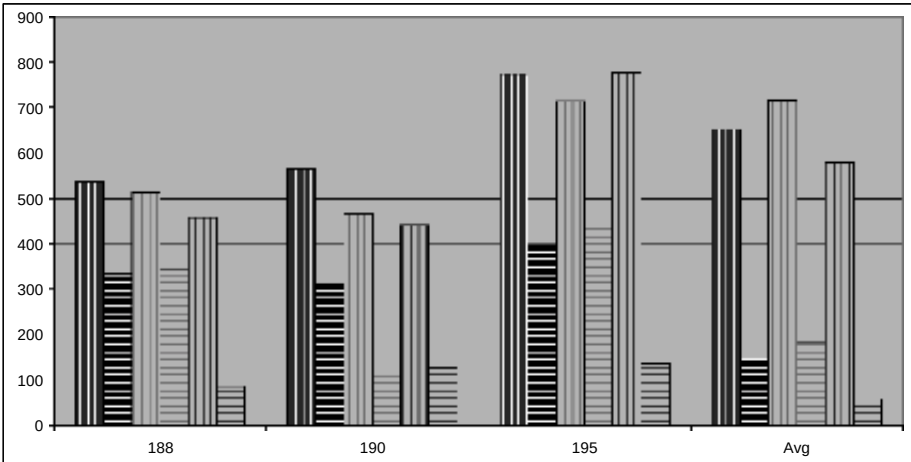


Fig. 5. “Easy topics”. Number of retrieved documents (vertical stripes) and number of relevant documents (horizontal stripes) for topics 188, 190, 195 compared to the average over all 25 topics. First 2 columns: German; second 2 columns: English; third 2 columns: Russian.

Whereas for the easy topics the relevant documents make up 51%, 56% and 22% of the retrieved documents per topic for English, German and Russian respectively, for the hard topics, only 7%, 8% or 0% of the documents were relevant for their respective languages (see Table 7 for average values).

One explanation for these differences might be the timeliness and specificity of topics. The topics yielding many relevant documents (188: Quality of life of elderly

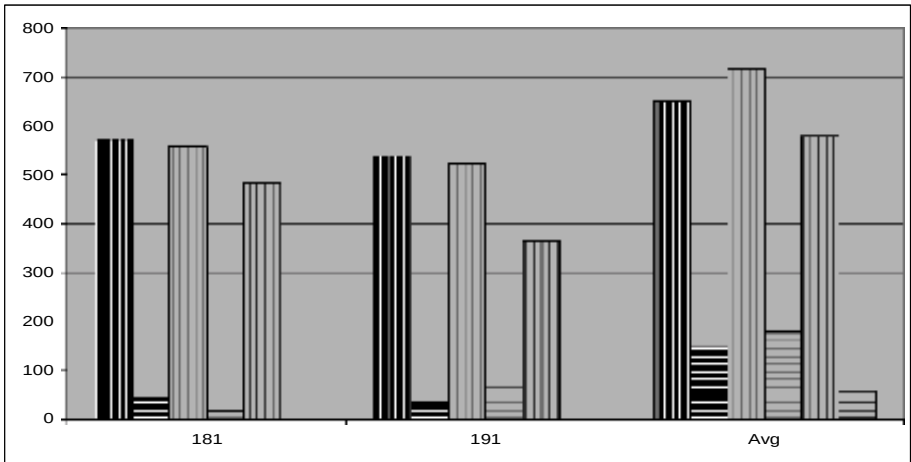


Fig. 6. “Hard topics”. Number of retrieved documents (vertical stripes) and number of relevant documents (horizontal stripes) for topics 181 and 191 compared to the average over all 25 topics. First 2 columns: German; second 2 columns: English; third 2 columns: Russian.

persons, 190: Mortality rate, 195: Integration policy) seem to be rather broad and ongoing themes in the social science literature. The other two topics (181: Torture in the constitutional state, 191: Economic elites in Eastern Europe and Russia) could be considered more specific and geared towards more recent time frames than others.

However, there are also topics, where one collection (language) seems to yield many results, whereas another collection will not produce many results. Figure 7 shows 2 examples.

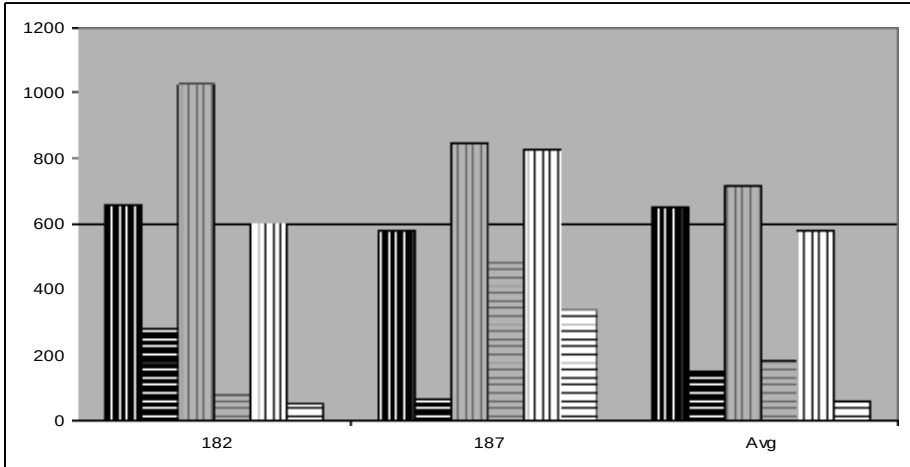


Fig. 7. Ambiguous topics. Number of retrieved documents (vertical stripes) and number of relevant documents (horizontal stripes) for topics 182 and 187 compared to the average over all 25 topics. First 2 columns: German; second 2 columns: English; third 2 columns: Russian.

Topic 182 (Family policy and national economy) finds many relevant documents in German, but few relevant English documents (even though many English documents were retrieved). Topic 187 (Economic elites in Eastern Europe and Russia) on the other hand finds many more relevant documents in the English and Russian collections than it does in the German. Whereas topic 182 is broader, topic 187 seems to be a bit more time-dependent.

Concluding from these examples, topic broadness and timeliness do not seem to be the only factors determining the success of a query in retrieval; however, they might be a start for more in-depth analysis.

4 Domain-Specific Experiments

Every group used the controlled vocabularies and structured data in some facility or other. One point of emphasis was query expansion with the help of the subject description provided by the thesauri. However, the translation and mapping tables were also used as bilingual dictionaries for cross-language experiments.

4.1 Retrieval Models

The Chemnitz group [10] used a redesigned version of their retrieval system based on the Lucene API and utilized two indices in retrieval: a structured index (taking the structure of the documents into account) and a plain index without considering the structure of the documents. To combine the two indices, a data fusion approach using the z-score introduced by the Unine group [12] was employed. They found that the unstructured indexed outperformed the structured one.

The Berkeley group [11] used a probabilistic model employing a logistic regression algorithm successfully used for cross-language retrieval since TREC-2 and implemented it with the Cheshire retrieval system.

Unine [3] used several retrieval models for comparison purposes: the classical tf idf vector space model, probabilistic retrieval with the Okapi algorithm and four variants of the DFR (Divergence from Randomness) approach as well as a language modelling approach. Data fusion was applied using the z-score to combine these different models. They also compared word-based and n-gram indexing for retrieval with the Russian language corpus.

The Xerox group [2] used a language modelling approach for their retrieval experiments.

4.2 Language Processing for Documents and Queries

Standard language processing for documents and queries in the form of stopword-removal and stemming or normalization was employed by all groups. The Unine group successfully developed a new light-weight stemmer for the Russian language.

For the German language, Unine and Xerox used a decompounding module to split German compounds whereas Berkeley and Chemnitz did not.

4.3 Query Expansion

Three of the groups focused on query expansion in some way or another. Berkeley used a version of Entry Vocabulary Indexes [4] based on the same logistic regression algorithm as their retrieval system to associate title and description terms from topics with controlled vocabulary terms from documents. Another approach was a thesaurus-lookup where title and description words were looked up in a thesaurus that combined all subject-describing keywords from the different collections. The terms from the controlled vocabularies were added to the query. As part of its standard retrieval process, the Cheshire system also implemented a blind feedback algorithm based on the Robertson and Sparck Jones term weights. Whereas the Entry Vocabulary Index approach worked better for the English target language, the thesaurus look-up worked better for German and Russian.

Unine used the Thesaurus for the Social Sciences to enhance queries with terms from the thesaurus. Thesaurus entries were indexed as documents and retrieved in response to query terms, then simply added to the query. They also used blind query feedback with Rocchio's formula as well as an idf-based approach described in [1]. The blind feedback approach improved the average precision of results, whereas the thesaurus expansion did not.

Xerox used lexical entailment to provide query expansion whereby a language modelling approach is employed to find similar terms from corpus documents in relation to query terms. They found that this approach outperformed simple blind feedback but a combined approach worked best.

4.4 Translation

Another focus of research was query translation, where the provided mapping tables were utilized as bilingual dictionaries.

Berkeley used the commercially available LEC Power Translator program for translation in all languages.

Chemnitz implemented a translation-plugin to their Lucene retrieval system utilizing well-known freely-available translation services like Babel Fish, Google Translate, PROMT and Reverso. They also used the bilingual mapping table from the thesauri for translation.

Finally, Xerox compared their Statistical Machine Translation System MATRAX with a sophisticated language-model-based approach of dictionary adaptation. Dictionary adaptation attempts to select one out of several translation possibilities for a term using a bilingual dictionary and calculating the probability of a target term given the language context of the source query term. They found that this approach worked well compared to the statistical machine translation system tested.

5 Results

In Appendix C of the Working Notes (http://clef-campaign.org/2007/working_notes/CLEF2007WN-Contents.html), mean average precision numbers (MAP) and recall-precision graphs for each run per task are listed.

6 Outlook

This year's experiments have shown that leveraging a controlled vocabulary for query expansion or translation can improve results in structured test collections. A new collection and new vocabulary (CSA Sociological Abstracts) was added and a mapping table between the CSA Thesaurus and the GIRT Thesaurus provided for experiments. As new collections are added and distributed search across several collections becomes more common, the seamless switching between controlled vocabularies becomes crucial to utilize expansion and translation techniques developed for individual collections.

For this purpose, several resources for terminology mapping have been developed at the German Social Science Information Centre [9]. Among them are over 40 bidirectional mappings between various controlled vocabularies. A web service to retrieve mapped terms is being developed. Besides the expansion of test collections, these vocabulary mapping services could be a future branch of research for the domain-specific track within CLEF.

Furthermore, the numbers of submitted runs against Russian as well as multilingual runs increased compared to previous years. Russian retrieval will remain an area of focus for this track. More translation tables (in English, German and Russian) for

inter-thesaurus switching as well as mappings between the GIRT thesaurus and the Russian INION thesaurus will be provided for further experiments.

Acknowledgements. We would like to thank Cambridge Scientific Abstracts for providing the documents for the new Sociological Abstracts test collection.

We greatly acknowledge the support of Natalia Loukachevitch and her colleagues from the Research Computing Center of M.V. Lomonosov Moscow State University in translating the topics into Russian as well as providing parts of the Russian relevance assessments.

Very special thanks also to Giorgio Di Nunzio and Nicola Ferro from the Information Management Systems (IMS) Research Group at the University of Padova for providing the DIRECT system and all their help in the assessment process and for providing the graphs and numbers for the results analysis.

Claudia Henning and Jeof Spiro did the German and English assessments; Monika Gonser and Oksana Schäfer provided the rest of the Russian assessments.

References

1. Abdou, S., Savoy, J.: Searching in Medline: Stemming, query expansion, and manual indexing evaluation. *Information Processing & Management* (to appear, 2007)
2. Clinchant, S., Renders, J.-M.: Query Translation through Dictionary Adaptation. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 182–187. Springer, Heidelberg (2008)
3. Fautsch, C., Dolamic, L., Abdou, S., Savoy, J.: Domain-Specific IR for German, English and Russian Languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 196–199. Springer, Heidelberg (2008)
4. Gey, F., Buckland, M., Chen, A., Larson, R.: Entry vocabulary – a technology to enhance digital search. In: *Proceedings of HLT2001, First International Conference on Human Language Technology*, San Diego, March 2001, pp. 91–95 (2001)
5. Girt Description: GIRT - Mono- and Cross-language Domain-Specific Information Retrieval (GIRT4) (2007), http://www.gesis.org/en/research/information_technology/girt4.htm
6. Kluck, M., Gey, F.: The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, pp. 48–56. Springer, Heidelberg (2001)
7. 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)
8. Kluck, M., Stempfhuber, M.: Domain-Specific Track CLEF 2005: Overview of Results and Approaches, Remarks on the Assessment Analysis. In: *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria, 21–23 September (2005), http://www.clef-campaign.org/2005/working_notes/workingnotes2005/kluck05.pdf
9. KoMoHe Project Website: Competence Center Modeling and Treatment of Semantic Heterogeneity (2007), http://www.gesis.org/en/research/information_technology/komohe.htm
10. Kürsten, J., Wilhelm, T., Eibl, M.: The XTRIEVAL Framework in CLEF 2007: Lessons Learned. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)

11. Larson, R.: Experiments in Classification Clustering and Thesaurus Expansion for Domain Specific Cross-Language Retrieval. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 188–195. Springer, Heidelberg (2008)
12. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval. In: Working Notes for the CLEF 2004 Workshop, Bath, UK, 15-17 September (2004), http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/22.pdf
13. Stempfhuber, M., Baerisch, S.: Domain-Specific Track CLEF 2006: Overview of Results and Approaches, Remarks on the Assessment Analysis. In: Working Notes for the CLEF 2006 Workshop, Alicante, Spain, 20-22 September (2006), http://www.clef-campaign.org/2006/working_notes/workingnotes2006/stempfhuber0CLEF2006.pdf

The XTREIVAL Framework at CLEF 2007: Domain-Specific Track

Jens Kürsten, Thomas Wilhelm, and Maximilian Eibl

Chemnitz University of Technology
Faculty of Computer Science, Chair Computer Science and Media
Straße der Nationen 62
09107 Chemnitz, Germany
jens.kuersten@cs.tu-chemnitz.de,
thomas.wilhelm@s2000.tu-chemnitz.de,
eibl@informatik.tu-chemnitz.de

Abstract. This article describes the architecture and configuration of the XTREIVAL (eXtensible reTRIEval and EVALuation) framework. A first prototype is described in [1]. For CLEF 2007 a second prototype was implemented which was focused on the cross-language aspect. Runs for all subtasks of the *Domain-Specific track* were submitted. The performance of our submitted runs was on average compared to other participating groups. Additional experiments on the *Multilingual task* demonstrated substantial improvement.

Keywords: Evaluation, Cross-Language Retrieval, Data Fusion.

1 Introduction

The XTREIVAL framework is part of the project *sachsMedia - Cooperative Producing, Storage and Broadcasting for Local Television Stations*¹ at Chemnitz University of Technology. This project does research in two fields: automatic annotation and retrieval of audiovisual media on the one hand and distribution of audiovisual media via digital video broadcasting (DVB) and IPTV on the other hand. Our project partners are local TV stations in Saxony (a federal state in Germany). The XTREIVAL framework is a first result of the annotation and retrieval aspect of the project.

In order to enable local TV stations to cooperate, a common database is set up. Within this database raw, produced and broadcasted material is stored by every cooperating TV station. This material needs to be described as comprehensively as possible in order to be easily searchable. On the one hand, the description - or annotation - of the material is carried intellectually according to principles of formal documentation. Alternatively, sophisticated methods of multimedia retrieval like object recognition and automated speaker recognition will be implemented.

¹ Funded by the *BMBF (German Federal Ministry of Education and Research)* *InnoProfile* program of the Innovation Initiative *Entrepreneurial Regions*.

The outline of the paper is as follows: Section 2 describes the architecture and design of the XTRIEVAL framework. An overview about the experiment objectives, configurations and results is given in Sect. 3. Section 4 summarizes the results of submitted runs for the *Domain-Specific* track. The final Sect. 5 concludes the experiments and gives an outlook to future work.

2 The Retrieval and Evaluation Framework XTRIEVAL

The framework consists of three functional components, which are illustrated in Fig.1. These three major parts are: the indexer, the actual search engine and the evaluation toolkit.

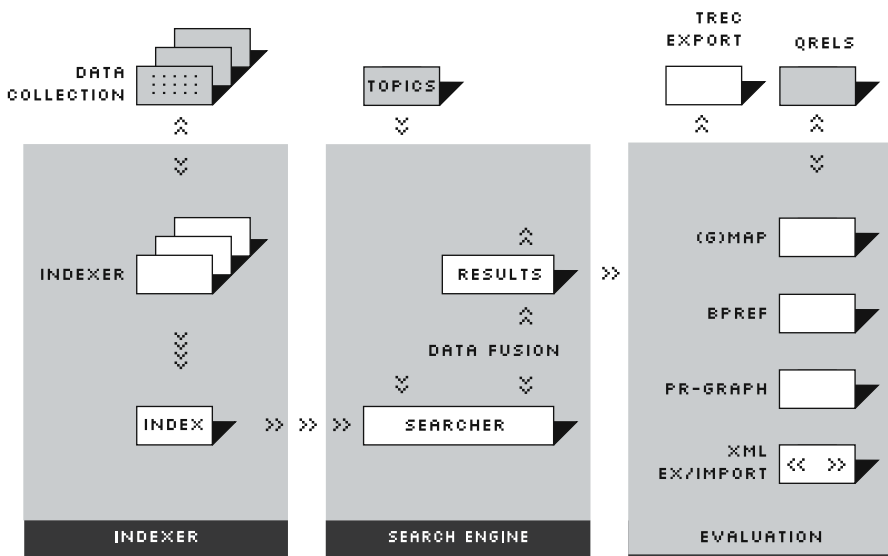


Fig. 1. XTRIEVAL framework architecture

The indexer is responsible for index creation. In the latest version of the system these indexes are created from different data collections (e.g. GIRT4 or IAPR TC-12). The implemented indexers are based on Lucene [2] and are inherited from an abstract indexer.

The search engine (SE) basically consists of a number of filters and a set of customizable search algorithms. The topic filters are entirely used to preprocess the queries in any necessary way (e.g. frequency analysis or translation). The configurable design of the SE enables an easy comparison of implemented search methods and information retrieval (IR) concepts like pseudo-relevance feedback. It is even possible to combine result lists retrieved either from different collections or from different retrieval models. In this case the final result list will be created

by using one of the well-known data fusion operators and a weighting for each of the combined result lists.

The evaluation component provides methods to compare the results obtained by the SE component. Here, several IR performance measures are offered: including precision-recall graph, mean average precision (MAP), geometric mean average precision (GMAP), precision at 20 (P20) and others. It is possible to save and load result lists for a later evaluation or comparison. Additionally, it is possible to import and export the results into the TREC format. To support the relevance judgements the import and export of QRELS is supported as well.

3 Experiment Objectives and Results

This year our group participates at the *Domain-Specific track* for the second time. Our first participation in 2006 focused on monolingual issues. This year we had two main objectives. First, we wanted to submit runs for all cross-lingual tasks. Therefore, we implemented a plug-in for the retrieval framework, which is capable to translate the topics to all languages of the track. The second objective was to investigate whether the combination of multiple index schemes could improve retrieval performance. Thus the XTREIVAL framework was designed to retrieve document lists from multiple indexes. Unfortunately, due to short implementation time and compatibility problems we had to abandon the local document clustering approach [3] which has led to very good results in 2006.

The general configuration of the system was as follows. We used a classic language processing chain to handle the topics, i.e. a stopword filter with the stopword lists provided by [4] and a stemming algorithm depending on the language (see the following subsections) as well as a standard tokenizer. A top-k pseudo-relevance feedback approach has been used to improve retrieval performance. We also used the term-frequency based topic preprocessor from last year to avoid searching for terms that are part of the query formulation and not the query itself.

3.1 Monolingual Runs

The monolingual experiments focused on comparing the performance of index schemes with the following differences: For each of the given data collections (DC) in English, German and Russian language, two types of indexes were created. The first index preserves the original structure by mapping all the data fields into corresponding fields in the index (i.e. 1-to-1 mapping for DC to index structure). The second index was created by mapping all elements - except for the *author* and the *doc-id* fields - of the DC structure (i.e. *title*, *abstract*, *controlled-term*, a.s.o.) into one field of the index, which corresponds to an n-to-1 mapping. For each language, document lists from both of those two indexes were retrieved by applying the z-score data fusion operator [5]. Table 1 summarizes the results of our experiments for the monolingual tasks.

² Not officially submitted experiment.

For the monolingual English task two different stemming algorithms were implemented: The Porter stemmer, obtained from the Snowball Project [6], and the Krovetz stemmer [7]. The aim was not to compare the performance of the different stemmers, but to combine their results during the retrieval step to improve retrieval performance.

For the monolingual German subtask, the German2 stemmer from the Snowball Project [6] was used. Due to limited implementation time neither we were able to adapt our decompounding algorithm from 2006 nor we could use thesauri for query expansion. But an additional experiment was conducted, where a collocation approach for controlled vocabulary was adapted, which is described in [8] and had been used successfully in experiments [9] for this task two years ago.

The experiments for the monolingual Russian task were based on a Russian analyzer and stemmer, which are part of an outdated version of the Lucene API [2]. Due to some errors in the configuration of the retrieval stage, the monolingual Russian experiments had to be repeated after the submission deadline.

Table 1. Domain-Specific monolingual retrieval performance

<i>identifier</i>	<i>language (corpora)</i>	<i>mapping: DC-to-index</i>	<i>MAP</i>	<i>GMAP</i>
cut_ds_mono_en_struct	EN (GIRT4, CSA)	1-to-1	0.1850	0.1124
cut_ds_mono_en_unstruct	EN (GIRT4, CSA)	n-to-1	0.2952	0.2208
cut_ds_mono_en_merged	EN (GIRT4, CSA)	merged	0.2985	0.2218
cut_ds_mono_de_struct	DE (GIRT4)	1-to-1	0.2631	0.1687
cut_ds_mono_de_unstruct	DE (GIRT4)	n-to-1	0.2887	0.2192
cut_ds_mono_de_merged	DE (GIRT4)	merged	0.2991	0.2189
cut_ds_mono_de_merged_add ²	DE (GIRT4)	merged	0.3495	0.2854
cut_ds_mono_ru_struct	RU (ISISS)	1-to-1	0.0898	0.0098
cut_ds_mono_ru_unstruct	RU (ISISS)	n-to-1	0.1293	0.0108
cut_ds_mono_ru_merged	RU (ISISS)	merged	0.1312	0.0119
cut_ds_mono_ru_merged_add ²	RU (ISISS)	merged	0.1523	-

The performance analysis of our monolingual experiments allows to draw the following conclusions. For the three languages and their corresponding corpora the n-to-1 collection-index mapping significantly outperformed the 1-to-1 mapping approach. This is due to the weighting of the 1-to-1 approach, where a fixed value is assigned to each field of the data collection. Additionally the merged run for each language outperformed the corresponding n-to-1 mapping experiment. That fact does not only manifest the n-to-1 mapping is superior to the 1-to-1 approach, but it also suggests not to abandon the 1-to-1 mapping. For that reason we will try to implement a better weighting scheme for the data fields of the collections, e.g. adaptive weighting (depending on the query terms).

3.2 Bilingual Runs

In order to conduct cross-lingual retrieval a translation plug-in for XTRIEVAL was developed to access on-line translation services. Namely, Google Translate

[10] and PROMT [11] had been used to receive the translation, because they performed best in some preliminary runs. The translation strategy was as follows: we assume that the topic titles are short and do not contain any sophisticated grammar constructs. Thus the topic titles were translated as phrases instead of a simple term by term translation, which was used for translating the topic descriptions. The grammar in the formulation of the topic description is assumed to cause problems with the machine translation systems mentioned above.

Additionally, the provided bilingual thesauri were used to support and compare the on-line translation process. That means that all terms of the topic title and description were looked up in the corresponding bilingual thesauri and when a term and its translation was found, it was appended to the translated query in a matter of a query expansion. The baseline configuration of each of our bilingual runs is the merged configuration of the corresponding monolingual run (see Sect. 3.1).

Table 2. Domain-Specific bilingual retrieval performance

<i>identifier</i>	<i>language pair</i>	<i>used bilingual thesaurus</i>	<i>MAP</i>	<i>GMAP</i>
cut_ds_mono_en_merged	EN	-	0.2985	0.2218
cut_ds_bili_ru2en_merged	RU-EN	no	0.2646 (-12.36%)	0.1502
cut_ds_bili_de2en_merged	DE-EN	no	0.1988 (-33.40%)	0.1453
cut_ds_bili_de2en_merged_thes	DE-EN	yes	0.2027 (-32.10%)	0.1504
cut_ds_mono_de_merged	DE	-	0.2991	0.2189
cut_ds_bili_ru2de_merged	RU-DE	no	0.1883 (-37.04%)	0.0327
cut_ds_bili_ru2de_merged_thes	RU-DE	yes	0.2047 (-31.56%)	0.0388
cut_ds_bili_en2de_merged	EN-DE	no	0.2012 (-32.73%)	0.0984
cut_ds_bili_en2de_merged_thes	EN-DE	yes	0.2721 (-09.03%)	0.1601
cut_ds_mono_ru_merged_add ²	RU	-	0.1523	-
cut_ds_bili_de2ru_merged	DE-RU	no	0.0938 (-28.51%)	0.0091
cut_ds_bili_de2ru_merged_thes	DE-RU	yes	0.0935 (-28.74%)	0.0092
cut_ds_bili_en2ru_merged	EN-RU	no	0.1142 (-25.02%)	0.0177
cut_ds_bili_en2ru_merged_add ²	EN-RU	no	0.1247 (-18.12%)	-

The results of our submissions for the bilingual task are shown in Tab. 2. The values in parentheses in the MAP column represent the percentile loss in retrieval performance compared to our best monolingual run. The last row of Tab. 2 represents the retrieval performance of an additional run, which results from the changes in the monolingual Russian runs from Sect. 3.1.

The performance of the submitted bilingual experiments allows to draw the general conclusion that the given bilingual thesauri increase retrieval performance - or do not decline it at least. The reason for that could be the missing or wrong translation of domain-specific terms returned by the on-line machine translation systems. Another interesting outcome is, that the English-Russian translation pair significantly outperforms the other language pairs for the German and the Russian target collection.

3.3 Multilingual Runs

For the submission of our multilingual runs the on-line translation plug-in (see Sect. 3.2) was also used. A special topic filter was implemented to translate the queries into the target languages. Then a special multi-index searcher retrieves the documents from the indexes of the corresponding data collections and merges the resulting documents into a single result list. A language mapping between the topic filter and the search algorithm was used to identify which language is appropriate for which data collection.

After a deep analysis of the multilingual runs, an error in the language mapping was found. Due to this error we had to repeat the multilingual experiments. Table 3 shows the retrieval performance of these additional runs. We also changed the data fusion to z-score merging instead of using the CMB-SUM operator [12] for those additional runs.

Table 3. Domain-Specific multilingual retrieval performance

<i>identifier</i>	<i>source language</i>	data fusion operator	<i>MAP</i>	<i>GMAP</i>
cut_ds_multi_en2x_merged	EN	CMB-SUM	0.0833	0.0399
cut_ds_multi_de2x_merged	DE	CMB-SUM	0.0842	0.0494
cut_ds_multi_ru2x_merged	RU	CMB-SUM	0.0508	0.0080
cut_ds_multi_en2x_merged_add ²	EN	z-score	0.1950	0.1285
cut_ds_multi_de2x_merged_add ²	DE	z-score	0.1756	0.1399
cut_ds_multi_ru2x_merged_add ²	RU	z-score	0.1346	0.0642

The significant increase in performance of the additional runs is due to the fact, that the result list from each of the used indexes were combined correctly, while in our submitted experiments only two result lists or even only one result list was used for the creation of the final result list. The differences in performance of the three source languages correlates to the best performing experiments of the monolingual runs, where German and English runs returned similar results and Russian experiments performed significantly worse. We assume this is due to the different data collections and their structure.

4 Result Analysis - Summary

The following list provides a summary of the analysis of our retrieval experiments for the *Domain-Specific track* at CLEF 2007:

- *Monolingual*: Compared to the performance of all submitted runs (also the ones of other groups), the results here performed rather on average. The results using the n-to-1 index mapping lead to a better performance in all languages. The query expansion approach based on term collocation analysis in the additional experiment on the German target collection significantly increased retrieval performance.

- *Bilingual*: The performance was about 10 to 18% worse than the performance of the monolingual tasks. This is assumed to be only a slight decline. Especially the combination of on-line translations and bilingual thesauri performed well.
- *Multilingual*: Here the results of our additional runs were astonishingly good. In our experiments with the data fusion approaches the z-score operator clearly outperformed the CMB-SUM operator.

5 Conclusion and Future Work

Due to the major changes in the XTRIEVAL framework, the performance of our experiments for the *Domain-Specific track* was not as good as expected. At the moment we plan further enhancements of our framework by integrating other retrieval frameworks into it (e.g. Lemur toolkit [13] or Terrier [14]). This is highly interesting for further investigation since these frameworks are based on different retrieval models. Furthermore, we are planning the inclusion of further multimedia data like audio and video.

References

1. Wilhelm, T., Eibl, M.: ImageCLEF 2006 Experiments at the Chemnitz Technical University. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 739–743. Springer, Heidelberg (2007)
2. The Apache Software Foundation: Lucene (1998-2006). Retrieved August 10, 2006 from the World Wide Web, <http://lucene.apache.org>
3. Kürsten, J., Eibl, M.: Monolingual Retrieval Experiments with a Domain-Specific Document Corpus at the Chemnitz University of Technology. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 178–185. Springer, Heidelberg (2007)
4. University of Neuchâtel, IIUN - Computer Science Department (2007). CLEF and Multilingual information retrieval. Retrieved November 20, 2007, from IIUN Web site, <http://members.unine.ch/jacques.savoy/clef/index.html>
5. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval. In: Working Notes for the CLEF 2004 Workshop, Bath, UK, 15-17 September (2004) Retrieved November 20, 2007, from CLEF Web site, http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/22.pdf
6. Porter, M.: The Snowball Project (2001-2007). Retrieved November 20, 2007, from Snowball Web site, <http://snowball.tartarus.org>
7. Krovetz, B.: Viewing morphology as an inference process. In: Proceedings of the 16th Annual International ACM SIGIR Conference, pp. 191–202 (1993)
8. Plaunt, C., Norgard, B.A.: An Association Based Method for Automatic Indexing with a Controlled Vocabulary. Journal of the American Society for Information Science 49(10), 888–902 (1998)

9. Petras, V.: How One Word Can Make all the Difference - Using Subject Metadata for Automatic Query Expansion and Reformulation. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria, 21-23 September (2005) Retrieved November 20, 2007, from CLEF Web site, http://www.clef-campaign.org/2005/working_notes/workingnotes2005/petras05.pdf
10. Google. Google Translate BETA (2007). Retrieved November 20, 2007, from Google Web site, http://www.google.com/translate_t
11. PROMT, Ltd. PROMT online-translator (2003-2007). Retrieved November 20, 2007, from PROMT Web site, <http://www.online-translator.com/text.asp>
12. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proceedings of the 2nd Text Retrieval Conference (TREC2), Gaithersburg, MD, vol. 500-215, pp. 243-252. NIST Special Publication (1993)
13. The Lemur Project . The Lemur Toolkit, University of Massachusetts (2007). Retrieved November 20, 2007, from Lemur Project Web site, <http://www.lemurproject.org/>
14. Terrier. Terabyte Retriever, University of Glasgow (2007). Retrieved November 20, 2007, from Terrier Web site, <http://ir.dcs.gla.ac.uk/terrier/>

Query Translation through Dictionary Adaptation

Stephane Clinchant and Jean-Michel Renders

Xerox Research Centre Europe, 6 ch. de Maupertuis, 38240 Meylan, France

`FirstName.LastName@xrce.xerox.com`

<http://www.xrce.xerox.com>

Abstract. Our participation to CLEF07 (Domain-specific Track) was motivated this year by assessing several query translation and expansion strategies that we recently designed and developed. One line of research and development was to use our own Statistical Machine Translation system (called Matrax) and its intermediate outputs to perform query translation and disambiguation. Our idea was to benefit from Matrax' flexibility to output more than one plausible translations and to train its Language Model component on the CLEF07 target corpora. The second line of research consisted in designing algorithms to adapt an initial, general probabilistic dictionary to a particular pair (query, target corpus); this constitutes some extreme viewpoint on the "bilingual lexicon extraction and adaptation" topic. For this strategy, our main contributions lie in a pseudo-feedback algorithm and an EM-like optimisation algorithm that realize this adaptation. A third axis was to evaluate the potential impact of "Lexical Entailment" models in a cross-lingual framework, as they were only used in a monolingual setting up to now. Experimental results on CLEF-2007 corpora (domain-specific track) show that the dictionary adaptation mechanisms appear quite effective in the CLIR framework, exceeding in certain cases the performance of much more complex Machine Translation systems and even the performance of the monolingual baseline. In most cases also, Lexical Entailment models, used as query expansion mechanisms, turned out to be beneficial.

Keywords: Domain-specific IR, Lexicon Extraction, Query Translation and Disambiguation, Dictionary Adaptation.

1 Introduction

When adopting a dictionary-based approach for query translation, the first and naive use of a dictionary consists in using all translations — possibly weighted — of each query word. Albeit simple, this approach does not address the *polysemy* of words, in the sense that it considers equally all the possible meanings. It is also very frequent that, for several reasons, lexicons originating from standard dictionaries, parallel corpora or comparable corpora often give spurious, irrelevant translations. Note though that the retrieval process is a disambiguating process in itself, in that spurious translations are generally filtered out simply by the

fact that it is very unlikely that they co-occur with other translations. Several approaches [1,2,3,4] resolve the translation of query with the notion of *coherence*. Each query term has candidate translation terms and a co-occurrence statistics can be computed between all the candidate translation terms; then an optimisation algorithm is used to solve some maximum coherence problem (find, for each query word, the best possible translation that maximizes its coherence with the other translations). The idea is that the query defines a lexical field. The more likely a candidate belongs to the lexical field, the better it is for translation.

2 Main Contribution: Dictionary Adaptation

Our contribution goes one step further, by adopting a more extreme viewpoint: starting from probabilistic dictionaries extracted by means of extraction techniques cited above, we will try to modify them and to adapt them to a particular (source query, target corpus) pair. In other words, our goal is to find the most relevant translations (and their associated weights), given the particular context of the whole query and the target corpus. This methodology can be further exploited in a categorization or clustering framework, if an equivalent processing is performed on the document to be classified instead of the query. This approach corresponds to a generalization of multi-word (complex term) bilingual terminology extraction: indeed, we try to find the best translations of the entire query, considered as a multi-word expression. A key point of the approach is to be query-focused: we are not trying to extract a large-span bilingual lexicon, but rather we are looking for some on-line, adapted lexicon, taking into account the particular characteristics of the user needs and the target corpus.

In a nutshell, the dictionary adaptation method could be derived from a cross-lingual extension of the monolingual mixture model for pseudo-relevance feedback introduced in [5], that heavily uses the “*Language Modelling for Information Retrieval*” Framework. The basic principle is to derive a language model of the “relevant concept” in the target language. All technical details of the methods are given in [6] and also, with more explanations, at URL <http://www.smart-project.eu/files/D51.pdf>. Note that the same algorithm realizes both the query enrichment and the dictionary adaptation. Note also that the translation/adaptation is limited to the words of the query (w_s) if we adopt a simple maximum likelihood language model for the query (what is assumed in this work). Lastly, but importantly, the role of the initial (probabilistic), non-adapted dictionary lies in providing the algorithm with a good starting candidate solution for the EM-algorithm that realizes the adaptation.

3 Secondary Contribution : Use of SMT’s Intermediate Outputs for Query Translation

Let us now come back to the use of Statistical Machine Translation (SMT) Systems for Information Retrieval. We do not really need a syntactically correct

query translation as, in most retrieval models, word order is not taken into account. Still, SMT could be of some value in solving the Query Translation task for the following reasons. Firstly, we have some flexibility in the choice and the building of the parallel corpus that is the basis of the alignment models; in particular, it is rather easy to concatenate large, general parallel corpora (e.g. the “JRC-Acquis Communautaire Corpus”) with smaller, but more specialised parallel corpora (we can even artificially duplicate the latter, to give it more weights in the resulting translation probabilities). Secondly, we can easily use the target corpus as the training corpus for the Language Model (LM) component of the SMT system so that, eventually, it automatically selects the most plausible sets of translations; in other words, it naturally solves the translation ambiguity issue and the phrasal translation problem through the use of LM adapted to the target corpus. However, this LM encodes some unnecessary order information: actually, we are more interested in the probability distribution of a set of words, rather than a sequence of words. Ideally, we would like the SMT tool to relax the order constraint, in order to provide the retrieval engine (not a human!) a set of plausible, coherent word translations. Thirdly, we can use the lattice of translation candidates as the new query representation; in this sense, it also realizes some query enrichment (giving for instance all valid synonyms of a query words), keeping the ambiguity when necessary.

In practice, we have chosen to use MATRAX, our home-made non-contiguous-phrase-based SMT tool. The choice of the training corpora is of course case-dependent, so that these details are considered in the more detailed report [6] and at URL: <http://www.smart-project.eu/files/D51.pdf> (experimental sections). Let us just mention that the n -best outputs of MATRAX are concatenated to form a new word distribution (in the target language) associated to the query. MATRAX is just one component of the cross-lingual retrieval; other components such as query expansion and pseudo-feedback methods are critical components and how they can be combined with MATRAX outputs is, once again, reported in the experimental part of the detailed papers.

4 Lexical Entailment

We also extensively used “Lexical Entailment” methods in our experiments. We recall here some of its features, as it appeared that Lexical Entailment, combined with query translation methods we propose in this work, performs pretty good. In Information Retrieval, Lexical Entailment (LE) [7,8] is a query expansion mechanism (curiously inspired from a cross-lingual framework) that models the probability that one term entails another, in a monolingual framework. Lexical Entailment can be understood as a probabilistic term similarity or as a unigram language model associated to a word (rather than to a document or a query). Let u be a term in the corpus, then lexical entailment models compute a probability distribution over terms v of the corpus $P(v|u)$. These probabilities can be used in information retrieval models to enrich queries and/or documents and to give a similar effect to the use of a semantic thesaurus. Unlike the construction of

semantic thesauri, lexical entailment is purely automatic, by extracting statistical relationships from the considered corpus. In practice, a sparse representation of $P(v|u)$ is adopted, where we restrict v to be one of the N_{max} terms that are the closest from u using an Information Gain metric. We refer to [7] for all technical and practical details of the method.

5 Lessons Learnt from CLEF 07 (Domain-Specific Track)

We refer to the introductory paper in [9] for details of the task, the corpora, the available resources and the queries. Experimental results of the methods introduced in the previous section in the case of CLEF2006 and CLEF 2007 - Domain Specific Track could be found in [6] and at URL <http://www.smart-project.eu/files/D51.pdf>.

All our bilingual runs follow the same schema “query translation” followed by a monolingual search (most often with PRF or query expansion in the target language). For the first step – query translation –, we used either the Statistical Machine Translation system (MATRAX), or one initial “standard” dictionary adapted following the strategy described hereabove.

We fed MATRAX with the JRC-AC (Acquis Communautaire) Corpus for the alignment models, and with our GIRT / CSA corpora (in the target language) for the language models. In this way, we can expect to introduce some bias or adaptation to our target corpus in the translation process, as the Language Model component of Matrax will favour translation and disambiguation consistent with this corpus. In order to increase the recall of what can be obtained with MATRAX, we intentionally kept the TOP5 most plausible translations given by MATRAX and concatenated them to obtain the new query in the target language (this indeed significantly increased the performance of the retrieval).

In order to perform lexicon adaptation, the choice of the initial dictionary is crucial to the task. We used two initial dictionaries that were at our disposal: the first one, CsaGirt, has been extracted from the concatenation of the GIRT and CSA thesauri. The second dictionary was ELRAC, composed as described before. Hence, to benefit from both sources, the dictionaries were merged hierarchically : an entry of the dictionary is added to the other one, if this entry is not already present in the master dictionary.

The main lessons and conclusions are the following:

- The absence of dictionary adaptation (considering all weighted, non-transformed translation candidates) has the consequence that any subsequent traditional (monolingual) relevance feedback algorithm we tried did not boost the performance of the retrieval. As the translated query is already noisy, it is likely that expanding it makes it unstable since feedback terms are mixed with irrelevant terms issued by the naive translations.
- With dictionary adaptation, we gain in performance for all dictionaries and translation directions. We obtain a global improvement ranging from 3% to 10%, and a relative improvement from 10% to 50% and an average gain of 6% for both directions and both dictionaries.

- A general dictionary with the adequate adaptation mechanism and a sufficient cover can be used for a specialized corpus, without a huge loss compared to a domain specific dictionary. Of course, domain specific dictionaries work better but they require external resources, or comparable corpora to be extracted from, whereas general dictionaries are always more easily available.
- the adaptation algorithm seems to be very stable and robust to the number of feedback documents. One can also notice , that much of the gain can be obtained using only the top 10 documents. We believe the stability is due to the initialization of algorithm with the initial, seed dictionary, which makes only non-zero entries serves as training data.
- If no other query expansion (in the target language) is done beyond the lexical entailment model, Matrax offers the best results (but recall that Matrax is significantly harder and more time-consuming to train than our simple dictionary extraction and adaptation).
- However, it seems that, once we want to adopt more complex PRF techniques after translation, there is a substantial advantage to use our dictionary adaptation method that, presumably, gives less noisy translations.
- Consequently, the best absolute performance are obtained by combining (1) the hierarchical building of the initial dictionary (the order in the hierarchy is dependent of the source and target languages, (2) adapting this initial dictionary with the proposed algorithm and (3) performing a rather sophisticated (PRF+Lexical Entailment) query expansion/enrichment in the target language.
- Finally, there is still a deficiency when the corpus is the English corpus both for monolingual and bilingual runs: we believe this is due to the unbalanced nature of the documents. German documents are longer in average (109 terms in average compare to 45 for english) and, consequently, more reliable, because they most often contain the abstract field.

6 Conclusions

Our main goal in this work was to validate two query translation and disambiguation strategies. The first one relies on the use of our Statistical Machine Translation tool, especially taking benefit from its flexibility to output more than one plausible translations and to train its Language Model component on the CLEF07 target corpora. The second one relies on a pseudo-feedback adaptation mechanism that performs simultaneously dictionary adaptation and query expansion.

Experimental results on CLEF-2007 corpora (domain-specific track) show that the dictionary adaptation mechanisms appear quite effective in the CLIR framework, exceeding in certain cases the performance of much more complex Machine Translation systems and even the performance of the monolingual baseline. The pseudo-feedback adaptation method turns out to be robust to the number of feedback documents and relatively efficient since we do not need to extract co-occurrence statistics. It is also robust to the noise in feedback documents, contrary to several traditional monolingual feedback methods that decreased their

performances in our experiments. Lastly, it enables to use general dictionaries in domain specific context with almost as good performance as domain specific dictionaries.

Aknowledgments

This work was partly supported by the IST Programme of the European Community, under the SMART project, FP6-IST-2005-033917. The authors also want to thank Francois Pacull for his greatly appreciated help in applying the MATRAX tools in CLEF07 experiments.

References

1. Kraaij, W., Nie, J.Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. *Comput. Linguist.* 29(3), 381–419 (2003)
2. Liu, Y., Jin, R., Chai, J.Y.: A maximum coherence model for dictionary-based cross-language information retrieval. In: *SIGIR 2005: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 536–543. ACM Press, New York (2005)
3. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language information retrieval. In: *SIGIR 2005: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 520–527. ACM Press, New York (2005)
4. Gao, J., Nie, J.Y., Zhou, M.: Statistical query translation models for cross-language information retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)* 5(4), 323–359 (2006)
5. Zhai, C., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: *Proceedings of the CIKM Conference*, pp. 403–410. ACM, New York (2001)
6. Clinchant, S., Renders, J.-M.: Xrce's participation to clef 2007 - domain specific track. In: *Working Notes of CLEF 2007*. Available On-line on the CLEF Web Site (2007)
7. Clinchant, S., Goutte, C., Gaussier, É.: Lexical entailment for information retrieval. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) *ECIR 2006*. LNCS, vol. 3936, pp. 217–228. Springer, Heidelberg (2006)
8. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: *PASCAL Challenges Workshop for Recognizing Textual Entailment* (2005)
9. Petras, V., Baerisch, S., Stempfhuber, M.: The domain-specific track at CLEF 2007. In: *Working Notes of CLEF 2007*. Available On-line on the CLEF Web Site (2007)

Experiments in Classification Clustering and Thesaurus Expansion for Domain Specific Cross-Language Retrieval

Ray R. Larson

School of Information
University of California, Berkeley, USA
`ray@ischool.berkeley.edu`

Abstract. In this paper we will describe Berkeley's approach to the Domain Specific (DS) track for CLEF 2007. This year we are using forms of the *Entry Vocabulary Indexes* and Thesaurus expansion approaches used by Berkeley in 2005[7]. Despite the basic similarity of approach, we are using quite different implementations with different characteristics. We are not, however, using the tools for de-compounding for German. All of the runs submitted were performed using the Cheshire II system. This year Berkeley submitted a total of 24 runs, including one for each subtask of the DS track. These include 6 Monolingual runs for English, German, and Russian, 12 Bilingual runs (4 X2EN, 4 X2DE, and 4 X2RU), and 6 Multilingual runs (2 EN, 2 DE, and 2 RU).

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the CLEF 2007 Domain Specific track. Last year for this track we used a baseline approach using only text retrieval methods without query expansion or use of the Thesaurus. This year we have focused instead on query expansion using Entry Vocabulary Indexes(EVIs)[5,7], and thesaurus lookup of topic terms. We continue to use probabilistic IR methods based on logistic regression.

All of the submitted runs for this year's Domain Specific track used the Cheshire II system for indexing and retrieval. The "Classification Clustering" feature of the system was used to generate the EVIs used in query expansion. The original approach for Classification Clustering in search was described in and [6]. While the method has changed considerably in implementation, the basic approach is still the same: topic-rich elements extracted from individual records in the database (such as titles, classification codes, or subject headings) are merged based on a normalized version of a particular organizing element (usually the classification or subject headings), and each such *classification cluster* is treated as a single "document" containing the combined topic-rich elements of all the individual documents that have the same values of the organizing element.

This paper first describes the probabilistic retrieval methods used and the EVI creation and search approach (Section 3.3). We then discuss our submissions for

the various DS sub-tasks and the results obtained. Finally we present some analysis of the results, conclusions and discussion of future approaches to this track.

2 The Retrieval Algorithms

As in previous years we used a version of the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a variety of cross-language retrieval tasks[3]. The algorithm described here was also used in our GeoCLEF and ImageCLEFPhoto submissions. The basic formula is:

$$\begin{aligned}
 \log O(R|C, Q) &= \log \frac{p(R|C, Q)}{1 - p(R|C, Q)} = \log \frac{p(R|C, Q)}{p(\bar{R}|C, Q)} \\
 &= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \frac{qt f_i}{ql + 35} \\
 &\quad + c_2 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \log \frac{tf_i}{cl + 80} \\
 &\quad - c_3 * \frac{1}{\sqrt{|Q_c|} + 1} \sum_{i=1}^{|Q_c|} \log \frac{ct f_i}{N_t} \\
 &\quad + c_4 * |Q_c|
 \end{aligned}$$

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 ,

$p(\bar{R}|C, Q)$ 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,

$qt f_i$ is the within-query frequency of the i th matching term,

tf_i is the within-document frequency of the i th matching term,

$ct f_i$ 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 though 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 $qt f_i$ 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.

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 $c_4 = 0.0929$. In addition to the direct retrieval of documents using this algorithm, 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], and is discussed further in our GeoCLEF paper in this volume.

3 Approaches for Domain Specific Retrieval

In this section we describe the specific approaches taken for our submitted runs for the Domain Specific track. First we describe the database creation and the indexing and term extraction methods used, and then the search features we used for the submitted runs.

3.1 Database Creation

For the purposes of this research we combined the GIRT German/English thesaurus along with the English and Russian mappings for the CSASA and ISSS databases to produce a multilingual thesaurus where elements from each of the original sources, as well as transliterations and capitalizations and the conversion of all data to UTF-8 encoding (this was also performed on the databases themselves before indexing). An example entry from this thesaurus is shown below:

```
<entry>
<german>Absatz</german>
<german-caps>ABSATZ</german-caps>
<scope-note-de>nicht im Sinne von Vertrieb</scope-note-de>
<english-translation>sale</english-translation>
<german_utf8>Absatz</german_utf8>
<russian>

сбыт

</russian>
<translit>sbyt </translit>
<mapping>
  <original-term>Absatz</original-term>
  <mapped-term>Sales</mapped-term>
</mapping>
<mapping>
```

```
<original-term>sale</original-term>  
<mapped-term>Sales</mapped-term>  
</mapping>  
</entry>
```

Note that the spacing around the Russian cyrillic term was inserted in the paper formatting process and was not in the original data.

Because not all of the terms had mappings, or equivalent Russian terms those parts are not present for all of the thesaurus entries.

3.2 Indexing and Term Extraction

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.

This year we used the Entry Vocabulary Indexes (search term recommenders) that were used in somewhat different form by Berkeley in previous years (see [7]). We did not, however, perform very well compared to other systems in the track this year. Given the changes in the collections used (the addition of the CSASA English collection and elimination of the Russian SocioNet data), it is not possible to directly compare MAP or other evaluation measures across years.

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. This is another aspect of our indexing for this year's Domain Specific task that reduced our results relative to last year.

3.3 Entry Vocabulary Indexes

As noted above earliest versions of Entry Vocabulary Indexes were developed to facilitate automatic classification of library catalog records, and first used in searching in [6]. Those used a simple frequency-based probabilistic model in searching, but a primary feature was that the "Classification clusters" were treated as documents and the terms associated with top-ranked clusters were combined with the original query, in a method similar to "blind feedback", to provide an enhanced second stage of search.

Our later work with EVIs used a maximum likelihood weighting for each term (word or phrase) in each classification. This was the approach described in [5] and used for Cross-language Domain-Specific retrieval for CLEF 2005. One limitation of that approach is that the EVI can produce maximum likelihood estimates for only a single term at a time, and alternative approaches needed to be explored for combining terms (see [7] for the various approaches).

In "Classification Clustering" various topic-rich elements are extracted from individual records in the database (such as titles, classification codes, or subject headings) and are merged into single records based on a normalized version of a

particular organizing element (usually the classification or subject headings, e.g., one record is created for each *unique* classification or subject heading). Each of these *classification clusters* is treated as a single “document” with the combined topic-rich elements of all the full documents with the same value of the organizing element. Searching the “Classification Clusters” uses the TREC2 algorithm with blind feedback described above. We then take some number of the top-ranked terms to expand the query for searching the document collections.

Two separate EVIs were built for the databases in each target language. The first used the contents of the “CONTROLLED-TERM-??” (or “KEYWORD” for Russian) fields as the organizing element. The second EVI used the contents of the “CLASSIFICATION-??” fields. Both of these EVIs were used in query expansion. One significant problem was that some records included multiple controlled terms in a single field instead of as separate fields. This was particularly common for the Russian “KEYWORD” terms. For this year we just ignored this problem rather than attempting to fix it, but we will be examining the effects in our analysis of the results.

3.4 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 the LEC Power Translator PC-based program.

Because all of our submitted runs this year used some form of query expansion, each required a 2-phase search process. The first phase involved a search in the EVI or the merged thesaurus, and the second phase combined some of the results of first phase search with the original query and used the expanded query to search the collections in the target language.

EVI Searches. For the monolingual and bilingual EVI searches (all those indicated in Table 1 with “EVI” in the “Exp.” column) the first search phase used all terms included in the “title” and “desc” fields of the topics (or the translated version of these fields). These terms were searched using the TREC2 algorithm with blind feedback to obtain a ranked result of classification clusters from the EVIs. The main or “organizing term” phrases for the top-ranked two clusters from the results for the “CONTROLLED-TERM” EVI, and the single top-ranked result phrase for the “CLASSIFICATION” EVI were extracted for use in the second phase.

For example, Topic #190 was searched using “mortality rate : find information on mortality rates in individual european countries” and the two EVIs yielded the following terms: “child mortality : infant mortality : demography and human biology; demography (population studies)”.

For the second phase search the original query was searched using the initial title+desc from the topic using the “topic” index and the expansion terms were

Table 1. Submitted Domain Specific Runs

Run Name	Description	Exp.	MAP
Berk_M_DE_CC_p15	Monolingual German	EVI	0.3150
Berk_M_DE_TH_p7	Monolingual German	Thes	0.3199
Berk_M_EN_CC_p15	Monolingual English	EVI	0.2814
Berk_M_EN_TH_p7	Monolingual English	Thes	0.2733
Berk_M_RU_CC_p15	Monolingual Russian	EVI	0.1390
Berk_M_RU_TH_p7	Monolingual Russian	Thes	0.1401
Berk_B_DEEN_CC_p15	German $\Rightarrow$ English	EVI	0.1096
Berk_B_DEEN_TH_p7	German $\Rightarrow$ English	Thes	0.1043
Berk_B_DERU_CC_p15	German $\Rightarrow$ Russian	EVI	0.0269
Berk_B_DERU_TH_p7	German $\Rightarrow$ Russian	Thes	0.0285
Berk_B_ENDE_CC_p15	English $\Rightarrow$ German	EVI	0.2412
Berk_B_ENDE_TH_p7	English $\Rightarrow$ German	Thes	0.2514
Berk_B_ENRU_CC_p15	English $\Rightarrow$ Russian	EVI	0.1348
Berk_B_ENRU_TH_p7	English $\Rightarrow$ Russian	Thes	0.1341
Berk_B_RUDE_CC_p15	Russian $\Rightarrow$ German	EVI	0.1520
Berk_B_RUDE_TH_p7	Russian $\Rightarrow$ German	Thes	0.1501
Berk_B_RUEN_CC_p15	Russian $\Rightarrow$ English	EVI	0.1757
Berk_B_RUEN_TH_p7	Russian $\Rightarrow$ English	Thes	0.1701
BerkMUDEp15	Multiling. from German	EVI	0.0468
BerkMUDEThp7	Multiling. from German	Thes	0.0486
BerkMUENp15	Multiling. from English	EVI	0.0884
BerkMUENTHp7	Multiling. from English	Thes	0.0839
BerkMURUp15	Multiling. from Russian	EVI	0.0414
BerkMURUThp7	Multiling. from Russian	Thes	0.0400

searched in the “subject” index, these searches were merged using a weighted sum for items in both lists. The estimated probability of relevance is a weighted combination of the initial estimated probability of relevance for the subject search and the probability of relevance for the entire document. Formally this is:

$$P(R \mid Q, C_{new}) = (X * P(R \mid Q, C_{subj})) + ((1 - X) * P(R \mid Q, C_{doc})) \quad (1)$$

Where X is a “pivot value” between 0 and 1, and $P(R \mid Q, C_{new})$, $P(R \mid Q, C_{subj})$ and $P(R \mid Q, C_{doc})$ are the new weight, the original subject search weight, and document weight for a given query. We found that a pivot value of 0.15 was most effective for CLEF2006 data when combining EVI and search queries.

Thesaurus-Based Searches. The basic steps for the searched doing thesaurus lookup is the same for EVIs, but the search structure is different. For the first phase search the topic title is searched among the language-appropriate main terms of the thesaurus, and the description is searched among all terms in the thesaurus entry. These intermediate results are combined using the pivot merger method

described about with a pivot weight of 0.55. The top two results are used, and both the language-appropriate main term, and the appropriate mapping terms are used for the query expansion. In the second phase the full topic title and desc fields are searched as topics, and the thesaurus terms are also searched as topics. These searches are combined using the pivot merge with a pivot weight of 0.07.

For topic #190 the first part of the query (i.e., the topic title and desc terms) is the same as for the EVI searches, but the second part of the search uses the terms yielded by the thesaurus search: “mortality : Infant mortality” (only a single thesaurus entry was retrieved in the search).

For multilingual searches, we combined the various translations of the topic title and desc fields produced by the LEC Power Translator for each source language and searched those combined translations in each target language. The results for each language were merged based on the MINMAX normalized score for each resultset. Within each language the same approaches were used as for EVI and Thesaurus-based expansion of bilingual and monolingual searches.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for all of our submitted runs for English, German and Russian are shown in Table 1, the Recall-Precision curves for these runs are not shown due to space limitations, but may be seen in the “notebook” version of the paper available on the CLEF Web site.

Since our experiments were conducted using the same topics, database, translation tools, and basic combination approaches for both EVIs and Thesaurus-based expansion, we were hoping to find a clear benefit for one approach versus the other. Unfortunately, the results are not at all clear. While EVIs seem to best results when English is the target language, the opposite is true for German and Russian targets. As always our multilingual results are significantly lower than monolingual or bilingual results for a given source language, with the exception of German⇒Russian, which is the lowest MAP of any of the runs.

Analysis of the differences between the Classification Clustering and thesaurus approach showed no statistically significant difference in performance between the two for a given source/target language set.

It is worth noting that the approaches used in our submitted runs provided the best results when testing with 2006 data and topics. However, as we discovered after the 2007 qrels were made available, some simpler approaches worked as well or better than the more complex methods described above. For example a simplified version of English monolingual search using only the topic title and desc fields, and searching each of those in the topic and subject indexes, and merging the results using a pivot value of 0.15 obtained a MAP result of 0.2848, compared to the 0.2814 obtained in our best submitted monolingual run. Further simplification to individual index searches does not, however provide results approaching those of the pivot-merged results. We suspect that the range of MAP scores for the track is different from previous years, or else our results are much worse than we thought they would be with the 2007 databases and topics.

5 Conclusions

We cannot say, overall, how effective query expansion by EVI or Thesaurus are relative to other approaches for this task. We can assume that there is very little difference in the effectiveness of the two methods, and that both seem to perform better than simple single-index “bag of words” searches of the collection contents.

We plan to conduct further runs to test whether modifications and simplifications, as well as combinations, of the EVI and Thesaurus-based approaches will provide can provide improved performance for the Domain Specific tasks.

References

1. Chen, A.: Multilingual information retrieval using english and chinese queries. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001*, Darmstadt, Germany, September 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 decompounding. *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. Gey, F., Buckland, M., Chen, A., Larson, R.: Entry vocabulary – a technology to enhance digital search. In: *Proceedings of HLT 2001, First International Conference on Human Language Technology*, San Diego, March 2001, pp. 91–95 (2001)
6. Larson, R.R.: Evaluation of advanced retrieval techniques in an experimental online catalog. *Journal of the American Society for Information Science* 43(1), 34–53 (1992)
7. 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)

Domain-Specific IR for German, English and Russian Languages

Claire Fautsch, Ljiljana Dolamic, Samir Abdou, and Jacques Savoy

Computer Science Department, University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchatel, Switzerland

{Claire.Fautsch,Ljiljana.Dolamic,Samir.Abdou,Jacques.Savoy}@unine.ch

Abstract. In participating in this domain-specific track, our first objective is to propose and evaluate a light stemmer for the Russian language. Our second objective is to measure the relative merit of various search engines used for the German and to a lesser extent the English languages. To do so we evaluated the *tf · idf*, Okapi, IR models derived from the *Divergence from Randomness* (DFR) paradigm, and also a language model (LM). For the Russian language, we find that word-based indexing using our light stemming procedure results in better retrieval effectiveness than does the 4-gram indexing strategy (relative difference around 30%). Using the German corpus, we examine certain variations in retrieval effectiveness after applying the specialized thesaurus to automatically enlarge topic descriptions. In this case, the performance variations were relatively small and usually non significant.

1 Introduction

In the domain-specific retrieval task we access the GIRT (German Indexing and Retrieval Test database) corpus, composed of bibliographic records extracted from two social science sources. This collection has grown from 13,000 documents in 1996 to more than 150,000 in 2005 (a more complete description of this corpus and the main results of this track can be found in [1]).

The manually assigned keywords contained in scientific documents are of particular interest to us, especially given that they are extracted from a controlled vocabulary by librarians. Through using this vocabulary and the corresponding thesaurus we hope to automatically enlarge the submitted queries and therefore improve retrieval performance.

2 Indexing and Searching Strategies

In order to obtain higher MAP values, we considered certain probabilistic models, such as the Okapi (or BM25). As a second probabilistic approach, we implemented variants of the DFR [2] (*Divergence from Randomness*) paradigm. We also examined an approach based on a statistical language model (LM) [3], also known as a non-parametric probabilistic model (a precise definition of these IR

models may be found at [4]). For comparison purpose, we also added the classical $tf \cdot idf$ model (with cosine normalization).

To measure retrieval performance, we adopted mean average precision (MAP) computed by `trec.eval`, based on 25 queries for the German and English corpora, and 22 for the Russian language. In the following tables, the best performance under a given condition is shown in bold type.

Table 1 lists evaluation results obtained using the Russian collection, combined with medium (TD) or long query formulations (TDN), along with two different indexing strategies (word-based using a light stemmer (inflectional only) and n -gram [5] scheme). An analysis of this data shows that the DFR model is the best performing of the IR models. This data also shows that the word-based approach uses the best indexing strategy. Taking this strategy as a baseline, the average performance difference for a 4-gram indexing strategy is around 29.5% (with TD query formulation) or 25% (with TDN queries).

Table 1. Evaluation of the Russian Corpus (22 queries)

Query Indexing	Mean average precision			
	TD word+light	TD 4-gram	TDN word+light	TDN 4-gram
Okapi	0.1630	0.0917	0.2064	0.1277
DFR-GL2	0.1639	0.1264	0.2170	0.1498
DFR-I(n)B2	0.1775	0.1052	0.2062	0.1433
LM	0.1511	0.1246	0.1952	0.1672
<i>tf idf</i>	0.1188	0.0918	0.1380	0.1229

Evaluations done on the German and English GIRT corpora are depicted in Table 2. In this case, we compared two query formulations (TD vs. TDN) and automatically enlarged topic descriptions, using the GIRT thesaurus. To achieve this we considered each entry in the thesaurus as a document, and then for each query we retrieved the thesaurus entries. Given the relatively small number of retrieved entries, we simply added all of them to the query to form a new and enlarged one. Although certain terms occurring in the original query were repeated, the procedure added related terms in other cases. If for example the topic included the name “Deutschland”, our thesaurus-based query expansion procedure might add the related term “BRD” and “Bundesrepublik”. Thus, these two terms would usually be helpful in retrieving more pertinent articles.

The results shown in Table 2 indicate that the best performing IR approach was usually the DFR-I(n)B2 model. Enlarging the query with terms extracted from the thesaurus does not improve the MAP. Rather, the contrary tends to be true, for they slightly reduce retrieval performance. Moreover, performance differences between the TD and TDN query formulations seem to be around 11.3% (German corpus with a decompounding stage) or 6.2% (English collection).

Upon looking at some queries more carefully, we can see when and why our search strategy fails to place pertinent articles at the top of the returned list. For

Table 2. Evaluation of German and English Corpora (25 queries)

Language Query Indexing	Mean average precision				
	German TD word	German TD + thesaurus	German TDN word	English TD word	English TDN word
Okapi	0.2616	0.2610	0.2927	0.2549	0.2501
DFR-GL2	0.2608	0.2599	0.2905	0.2710	0.2852
DFR-I(n)B2	0.2898	0.2877	0.2983	0.3130	0.3254
LM	0.2526	0.2336	0.2993	0.2603	0.2929
<i>tf idf</i>	0.1835	0.1805	0.2019	0.1980	0.2091

the German corpus, using the GIRT thesaurus, our system automatically added the term “Osterweiterung” related to the query term “Europäisch”. In general a relationship exists between these two terms but not in the context of Topic #199 (“Europäische Klimapolitik”). Generally, specific search terms would not have an entry in the GIRT thesaurus, yet for more frequent and less important words we might find some related terms in the thesaurus. Adding such terms did not help us find more relevant items.

From our observations we noted that another source of failure was the use of different word phrases to express the same concept. For Topic #171 (“Sibling relations”) there were two relevant items using the term “семейные” (family) but not the word “братьями” (“brothers”) or “сестрами” (“sisters”) used in the Russian topic formulation. Finally our search system encountered a real problem with Topic #192 (“System change and family planning in East Germany”). In this case, the only term common to the query formulations and the single relevant article was the frequently appearing noun “Germany”

3 Official Results

To define our official runs as described in Table 3, we first applied a pseudo-relevance feedback using Rocchio’s formulation [6] with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add m terms extracted from the k best ranked documents (the exact values used in our experiments are listed in Table 3).

In a second step, we combined three or four probabilistic models, representing both the parametric (Okapi and DFR) and non-parametric(LM) approaches. All runs were fully automatic and in all cases we applied the same data fusion approach (Z-score [4]). For the German corpus however we applied our decomposing approach (denoted by “dec.” in the “Index” column). For the English corpus our data fusion strategy clearly enhanced retrieval performance, but for the German or Russian, we obtained only slight improvements.

For our participation in this domain-specific evaluation campaign, we proposed a new light stemmer for the Russian language. The resulting MAP (see Table 1) shows that for this Slavic language our approach may produce better MAP than a 4-gram approach (relative difference around 30%). For the German

Table 3. Description and MAP Results for Our Best Official Monolingual Runs

Language	Index	Query	Model	Query exp.	MAP	comb. MAP
German UniNEde3	dec.	TD	PL2	10 docs/120 terms	0.3383	Z-score
	dec.	TD	InB2		0.2898	0.3535
	dec.	TD	PL2	10 docs/120 terms	0.3431	
	dec.	TD	InB2	10 docs/230 terms	0.3444	
English UniNEen1	word	TD	GL2	10 docs/100 terms	0.3080	Z-score
	word	TD	PB2	10 docs/150 terms	0.3165	0.3472
	word	TD	InB2		0.3130	
Russian UniNERu3	word	TD	Okapi	5 docs/50 terms	0.1579	Z-score
	4-gram	TD	LM	5 docs/50 terms	0.1331	0.1648
	word	TD	LM	10 docs/60 terms	0.1645	(0.1450)
	4-gram	TD	GL2	5 docs/50 terms	0.1335	

corpus, we tried to exploit the specialized thesaurus in order to improve the resulting MAP, yet retrieval effectiveness differences are rather small. We thus believe that a more specific query enrichment procedure is needed, one that is able to take the various different term-term relationships into account, along with the occurrence frequencies for the potential new search terms. Upon comparing the various IR models (see Table 1), we found that the I(n)B2 model derived from the *Divergence from Randomness* (DFR) paradigm would usually provide the best performance.

Acknowledgments. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

1. Petras, V., Baerisch, S., Stempfhuber, M.: The Domain-Specific Track at CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 160–173. Springer, Heidelberg (2008)
2. 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)
3. Hiemstra, D.: Using Language Models for Information Retrieval. PhD Thesis (2000)
4. Dolamic, L., Savoy, J.: Stemming Approaches for East European Languages. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 37–44. Springer, Heidelberg (2008)
5. McNamee, P., Mayfield, J.: Character N-gram Tokenization for European Language Text Retrieval. *IR Journal* 7, 73–97 (2004)
6. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New Retrieval Approaches Using SMART. In: *Proceedings TREC-4*, Gaithersburg, pp. 25–48 (1996)

Overview of the CLEF 2007 Multilingual Question Answering Track

Danilo Giampiccolo¹, Pamela Forner¹, Jesús Herrera², Anselmo Peñas³,
Christelle Ayache⁴, Corina Forascu⁵, Valentin Jijkoun⁶, Petya Osenova⁷,
Paulo Rocha⁸, Bogdan Sacaleanu⁹, and Richard Sutcliffe¹⁰

¹ CELCT, Trento, Italy

{giampiccolo, forner}@celct.it

² Departamento de Ingeniería del Software e Inteligencia Artificial,
Universidad Complutense de Madrid, Spain

jesus.herrera@fdi.ucm.es

³ Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain
anselmo@lsi.uned.es

⁴ ELDA/ELRA, Paris, France

ayache@elda.fr

⁵ Faculty of Computer Science, University “Al. I. Cuza” of Iași, Romania Institute for
Computer Science, Romanian Academy, Iași, Romania

corinfor@info.uaic.ro

⁶ Informatics Institute, University of Amsterdam, The Netherlands
jijkoun@science.uva.nl

⁷ BTB, Bulgaria

petya@bultreebank.org

⁸ Linguatca, SINTEF ICT, Norway and Portugal

Paulo.Rocha@alfa.di.uminho.pt

⁹ DFKI, Germany

Bogdan.Sacaleanu@dfki.de

¹⁰ DLTC, University of Limerick, Ireland

richard.sutcliffe@ul.ie

Abstract. The fifth QA campaign at CLEF [1], having its first edition in 2003, offered not only a main task but an Answer Validation Exercise (AVE) [2], which continued last year’s pilot, and a new pilot: the Question Answering on Speech Transcripts (QAST) [3, 15]. The main task was characterized by the focus on cross-linguality, while covering as many European languages as possible. As novelty, some QA pairs were grouped in clusters. Every cluster was characterized by a topic (not given to participants). The questions from a cluster possibly contain co-references between one of them and the others. Finally, the need for searching answers in web formats was satisfied by introducing Wikipedia¹ as document corpus. The results and the analyses reported by the participants suggest that the introduction of Wikipedia and the topic related questions led to a drop in systems’ performance.

¹ <http://wikipedia.org>

1 Introduction

Inspired in previous TREC evaluation campaigns, QA tracks have been proposed at CLEF since 2003. During these years, the effort of the organizers has been focused on two main issues. One of them was to offer an evaluation exercise characterized by cross-linguality, covering as many languages as possible. From this perspective, major attention has been given to European languages, not only adding at least one new language every year, but maintaining the catalogue of offered ones, except for Finish, which only could be offered in the 2005 edition. 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 2007 a pilot task on *Question Answering on Speech Transcripts* and a subsidiary task on Answer Validation (AVE) were proposed together with a *main* task. As far as the latter is concerned, the most significant innovations were the introduction of topic-related questions and the possibility to search for answers in Wikipedia. The topic-related questions consisted of clusters of questions which were related to the same topic. The requirement for related questions on a topic necessarily implies that the questions will refer to common concepts and entities within the domain in question. This accomplished either by co-reference either by anaphoric reference to the topic declared implicitly in the first question or in its answer. As far as the other major innovation of this year’s campaign, beside the data collections composed of news articles provided by ELRA/ELDA, also Wikipedia was considered, capitalizing on the experience of the WiQA pilot task proposed in 2006.

As general remark, the positive trend in participation registered in the previous campaigns was inverted for first time in the history of the QA@CLEF.

As reflected in the results, the task proved to be more difficult than expected, as in comparison with last year’s results dropped both in the multi-lingual subtasks and in the monolingual subtasks.

QA@CLEF 2007 was carried out according to the spirit of the campaign, consolidated in previous years. Beside the classical main task, an *Answer Validation Exercise* [13] and a pilot task on *Question Answering on Speech Transcripts* [15] were proposed:

- the *main* task, divided into several monolingual and bi-lingual sub-tasks, is described in this paper.
- the *Answer Validation Exercise* (AVE) continued the successful experiment proposed in 2006. In this task, systems were required to emulate human assessment of QA responses and decide whether an *Answer* to a *Question* is correct or not according to a given *Text*. Results were evaluated against the QA human assessments [2]. The overview of this exercise can be found in this volume [13].
- the *Question Answering on Speech Transcripts* (QAST) pilot task aimed at providing a framework in which QA systems can be evaluated when the answers to factual and definition questions must be extracted from spontaneous speech transcriptions. The main goals of this pilot were:
 - comparing the performances of the systems dealing with both types of transcriptions

- measuring the loss of each system due to the state of the art of the Automatic Speech Recognition (ASR) technology.
- in general, motivating and driving the design of novel and robust factual QA architectures for automatic speech transcriptions [3]. The overview of this exercise can be found in this volume [15].

This paper describes the preparation process and presents the results of the QA track at CLEF 2007. In section 2, the tasks of the track are described in detail. The results are reported in section 3. In section 4, some final analysis about this campaign is given. And section 5 consists of a draft about what should be addressed in the near future of QA@CLEF.

2 Task Description

As far as the main task is concerned, the consolidated procedure was followed, although some relevant innovations were introduced.

Following the example of TREC, this year the exercise consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and possibly contained co-references between one question and the others. Neither the question types (F, D, L) nor the topics were given to the participants.

The systems were fed with a set of 200 questions -which could concern facts or events (F-actoid questions), definitions of people, things or organisations (D-efinition questions), or lists of people, objects or data (L-ist questions)- and were asked to return one exact answer, where *exact* meant that neither more nor less than the information required was given.

The answer needed to be supported by the docid of the document in which the exact answer was found, and by portion(s) of text, which provided enough context to support the correctness of the exact answer. Supporting texts could be taken from different sections of the relevant documents, and could sum up to a maximum of 700 bytes. 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 *in-exact*. As in previous years, the exact answer could be exactly copied and pasted from the document, even if it was grammatically incorrect (e.g.: inflectional case did not match the one required by the question). Anyway, systems were also allowed to use natural language generation in order to correct morpho-syntactical inconsistencies (e.g., in German, changing *dem Presidenten* into *der President* if the question implies that the answer is in nominative case), and to introduce grammatical and lexical changes (e.g., QUESTION: *What nationality is X?* TEXT: *X is from the Netherlands* → EXACT ANSWER: Dutch).

The subtasks 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.

Ten source languages were considered, namely, Bulgarian, Dutch, English, French, German, Indonesian, Italian, Portuguese, Romanian and Spanish. All these languages

Table 1. Tasks activated in 2007 (coloured cells)

TARGET LANGUAGES (corpus and answers)										
SOURCE LANGUAGES (questions)		BG	DE	EN	ES	FR	IT	NL	PT	RO
	BG									
	DE									
	EN									
	ES									
	FR									
	IN									
	IT									
	NL									
	PT									
	RO									

were also considered as target languages, except for Indonesian, which had no news collections available for the queries and, as was done in the previous campaigns, used the English question set translated into Indonesian (IN).

As shown in Table 1, 37 tasks were proposed:

- 8 Monolingual -i.e. Bulgarian (BG), German (DE), Spanish (ES), French (FR), Italian (IT), Dutch (NL), Portuguese (PT) and Romanian (RO);
- 29 Cross-lingual.

Anyway, as Table 2 shows, not all the proposed tasks were then carried out by the participants.

Table 2. Tasks chosen by at least 1 participant in QA@CLEF campaigns

	MONOLINGUAL	CROSS-LINGUAL
CLEF-2004	6	13
CLEF-2005	8	15
CLEF-2006	7	17
CLEF-2007	7	11

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. English was still both source and target language in the cross-language tasks.

2.1 Questions Grouped by Topic

The procedure followed to prepare the test set was much different from that used in the previous campaigns. First of all, each organizing group, responsible for a target language, freely chose a number of topics. For each topic, one to four questions were generated. Topics could be not only named entities or events, but also other categories such as objects, natural phenomena, etc. (e.g. George W. Bush; Olympic Games; notebooks; hurricanes; etc.). The set of ordered questions were related to the topic as follows:

- the topic was named either in the first question or in the first answer
- the following questions could contain co-references to the topic expressed in the first question/answer pair.

Topics were not given in the test set, but could be inferred from the first question/answer pair. For example, if the topic was *George W. Bush*, the cluster of questions related to it could have been:

Q1: *Who is George W. Bush?*; Q2: *When was he born?*; Q3: *Who is his wife?*

The requirement for questions related to a same topic necessarily implies that the questions refer to common concepts and entities within the domain. In a series of questions this is accomplished by co-reference – a well known phenomenon within Natural Language Processing which nevertheless has not been a major factor in the success of QA systems in previous CLEF workshops. The most common form is pronominal anaphoric reference to the topic declared in the first question, e.g.:

Q4: *What is a polygraph?*; Q5: *When was **it** invented?*

However, other forms of co-reference occurred in the questions. Here is an example:

Q6: *Who wrote the song "Dancing Queen"?*; Q7: *How many people were in **the group**?*

Here *the group* refers to an entity expressed not in the question but only in the answer. However the QA system does not know this and has to infer it, a task which can be very complex, especially if the topic is not provided in the test set.

2.2 Addition of Wikipedia

Another major innovation of this year's campaign concerned the corpora at which the questions were aimed. In fact, beside the data collections composed of news articles provided by ELRA/ELDA (see Table 3), also Wikipedia was considered, capitalizing on the experience of the WiQA pilot task proposed in 2006 [9].

The Wikipedia pages in the target languages, as found in the version of November 2006, could be used. Romanian, which was addressed as a target language for the first

Table 3. Document collections used in QA@CLEF 2007

TARGET LANG.	COLLECTION	PERIOD	SIZE
[BG] Bulgarian	Sega	2002	120 MB (33,356 docs)
	Standart	2002	93 MB (35,839 docs)
[DE] German	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)
[EN] English	Los Angeles Times	1994	425 MB (113,005 docs)
	Glasgow Herald	1995	154 MB (56,472 docs)
[ES] Spanish	EFE	1994	509 MB (215,738 docs)
	EFE	1995	577 MB (238,307 docs)
[FR] French	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)
[IT] Italian	La Stampa	1994	193 MB (58,051 docs)
	Italian SDA	1994	85 MB (50,527 docs)
	Italian SDA	1995	85 MB (50,527 docs)
[NL] Dutch	NRC Handelsblad	1994/1995	299 MB (84,121 docs)
	Algemeen Dagblad	1994/1995	241 MB (106,483 docs)
[PT] Portuguese	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)

time, had Wikipedia² as the only document collection, because there was no newswire Romanian corpus. The “snapshots” of Wikipedia were made available for download both in XML and HTML versions. The answers to the questions had to be taken from actual entries or articles of Wikipedia pages. Other types of data such as images, discussions, categories, templates, revision histories, as well as any files with user information and meta-information pages, had to be excluded.

One of the major reasons for using Wikipedia was to make a first step towards web formatted corpora where to search for answers.

As nowadays so large information sources are available on the web, this is may be considered a desirable next level in the evolution of QA systems. An important advantage of Wikipedia is that it is freely available for all languages so far considered. Anyway the variation in size of Wikipedia, depending on the language, is still problematic.

² http://static.wikipedia.org/downloads/November_2006/ro/

2.3 Types of Questions

As far as the question types are concerned, as in previous campaigns, the three following categories were considered:

1. *Factoid questions*, fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc. We consider the following 8 answer types for factoids:
 - PERSON, e.g.: Q8: *Who was called the “Iron-Chancellor”?* A8: *Otto von Bismarck.*
 - TIME, e.g.: Q9: *What year was Martin Luther King murdered?* A9: *1968.*
 - LOCATION, e.g.: Q10: *Which town was Wolfgang Amadeus Mozart born in?* A10: *Salzburg.*
 - ORGANIZATION, e.g.: Q11: *What party does Tony Blair belong to?* A11: *Labour Party.*
 - MEASURE, e.g.: Q12: *How high is Kanchenjunga?* A12: *8598m.*
 - COUNT, e.g.: Q13: *How many people died during the Terror of PoPot?* A13: *1 million.*
 - OBJECT, e.g.: Q14: *What does magma consist of?* A14: *Molten rock.*
 - OTHER, i.e. everything that does not fit into the other categories above, e.g.: Q15: *Which treaty was signed in 1979?* A15: *Israel-Egyptian peace treaty.*
2. *Definition questions*, questions such as “What/Who is X?”, and are divided into the following subtypes:
 - PERSON, i.e., questions asking for the role/job/important information about someone, e.g.: Q16: *Who is Robert Altmann?* A16: *Film maker*
 - ORGANIZATION, i.e., questions asking for the mission/full name/important information about an organization, e.g.: Q17: *What is the Knesset?* A17: *Parliament of Israel.*
 - OBJECT, i.e., questions asking for the description/function of objects, e.g.: Q18: *What is Atlantis?* A18: *Space Shuttle.*
 - OTHER, i.e., question asking for the description of natural phenomena, technologies, legal procedures etc., e.g.: Q19: *What is Eurovision?* A19: *Song contest.*
3. *closed list questions*: i.e., questions that require one answer containing a determined number of items, e.g.: Q20: *Name all the airports in London, England.* A20: *Gatwick, Stansted, Heathrow, Luton and City.*

As only one answer was allowed, all the items had to be present in sequence in the document and copied, one next to the other, in the answer slot.

Besides, all types of questions could contain a temporal restriction, i.e. a temporal specification that provided important information for the retrieval of the correct answer, for example:

Q21: *Who was the Chancellor of Germany from 1974 to 1982?*
 A21: *Helmut Schmidt.*

Q22: Which book was published by George Orwell in 1945?

A22: Animal Farm.

Q23: Which organization did Shimon Perez chair after Isaac Rabin's death?

A23: Labour Party Central Committee.

Some questions could have no answer in the document collection, and in that case the exact answer was "NIL" and the answer and support docid fields were left empty. A question was assumed to have no right answer when neither human assessors nor participating systems could find one.

The distribution of the questions among these categories is described in Table 4. Each question set was then translated into English, which worked as inter-language during the translation of the datasets into the other tongues for the activated cross-lingual subtasks.

Table 4. Test set breakdown according to question type, number of participants and number of runs

	F	D	L	T	NIL	# Participants	# Runs
BG	158	32	10	12	0	0	0
DE	164	28	8	27	0	2	5
EN	161	30	9	3	0	5	8
ES	148	42	10	40	21	5	5
FR	148	42	10	40	20	2	2
IT	147	41	12	38	20	1	1
NL	147	40	13	30	20	0	0
PT	143	47	9	23	18	6	8
RO	160	30	10	52	7	1	2

2.4 Formats

As the format is concerned, this year both input and output files were formatted as an XML file. For example, the first three questions in the EN-DE test set, i.e. English questions that hit a German document collection - were represented as follows:

```
- <input>
<q target="DE" source="EN" id="0001" group_id="3600">Who is the singer
of the band U2?</q>
<q target="DE" source="EN" id="0002" group_id="3601">What age did El-
vis Presley die?</q>
<q target="DE" source="EN" id="0003" group_id="3602">Where is the
Statue of Liberty located?</q>
```

An example of system output which answered the above questions was the following:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
- <output>
- <a score="1.4222687" run_id="dfki071ende" group_id="3600" q_id="0001">
<answer>Annett Louisan</answer>
- <support>
<s_id>de_xml/p01/161699.xml</s_id>
```

```

<s_string>Aber auch Sänger und Sängerinnen wie Kate Ryan, Annett Loui-
san, die Band Overground, die Band Sportfreunde Stiller, Xavier
Naidoo, Die Fantastischen Vier, das DJ-Duo Blank & Jones und
Melendiz nutzten die U-Bahn und Kate Ryan Bahnhöfe für Kate Ryan
Musikvideos.</s_string>
</support>
</a>
- <a score="14.129519" run_id="dfki071ende" group_id="3601" q_id="0002">
<answer>42</answer>
- <support>
<s_id>SDA.950109.0187.xml</s_id>
<s_string>Elvis Presley war 1977 im Alter von 42 Jahren gestor-
ben.</s_string>
</support>
</a>
- <a score="0.009016844" run_id="dfki071ende" group_id="3602"
q_id="0003">
<answer>USA</answer>
- <support>
<s_id>de_xml/p09/984837.xml</s_id>
<s_string>Henry Martin war ein Freimaurer; zusammen mit Edouard René
Lefèvre de Laboulaye, den Enkeln des Marquis de Lafayette und
Frédéric Auguste Bartholdi, einem jungen Künstler aus dem Elsaß,
war Henry Martin maßgeblich an der Finanzierung der Frei-
heitsstatue, einem Geschenk an die USA beteiligt, deren Einweihung
Henry Martin nicht mehr miterlebte.</s_string>
</support>
</a>

```

2.5 Evaluation

As far the evaluation process is concerned, no changes were made with respect to the 2006 edition. Human judges assessed the exact answer (i.e. the shortest string of words which is supposed to provide the exact amount of information to answer the question) 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.

As regards the evaluation measures, the main one was accuracy, defined as the average of $SCORE(q)$ over all 200 questions q , where $SCORE(q)$ is 1 in the first answer to q in the submission file is assessed as R, and 0 otherwise.

In addition most assessor groups computed the following measures:

- K1 [6];
- Confident Weighted Score (CWS) [17].

3 Results

As far as accuracy is concerned, scores were generally far lower this year than usual, as Figure 1 shows. Although comparison between different languages and years is not

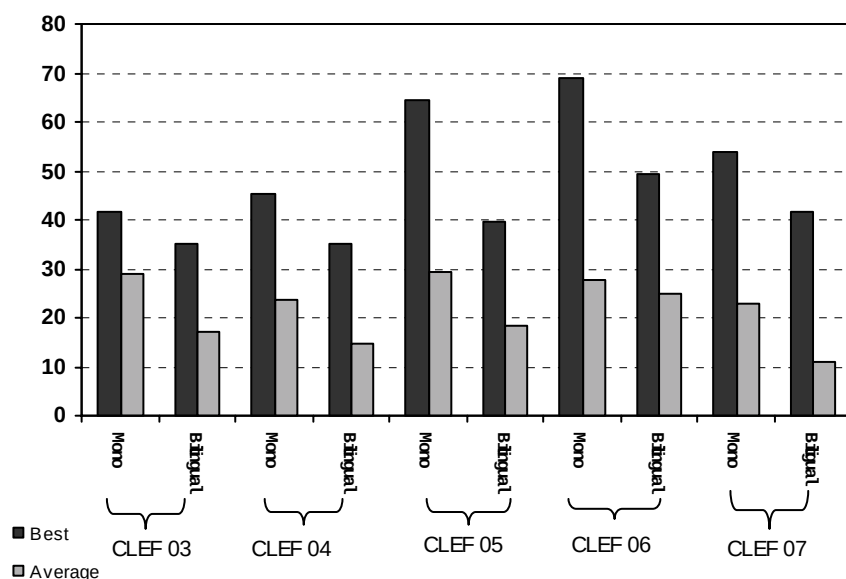


Fig. 1. Best and average scores in QA@CLEF campaigns

possible, we can observe in Figure 1 some trends this year: best accuracy both in the monolingual and the bilingual tasks decreased considerably.

This is also true for average performances. This year a neat decrease has been recorded in the bilingual tasks, due also to the presence of systems which participated for the first time, achieving very low score in tasks which are quite difficult also for veterans.

As a general remark, it can be said that the new factors introduced this year appear to have had an impact on the performances of the systems. As more than one participant has noticed, there has been not enough time to adjust the systems to the new requirements.

3.1 Participation

After years of constant growth, the number of participants has decreased in 2007 (see Table 5) due to the new challenges introduced in the exercise.

The geographical distribution has anyway remained almost the same, recording a new entry of a group from Australia. No participants took part to any Bulgarian tasks.

Table 5. Number of participants in QA@CLEF

	America	Europe	Asia	Australia	TOTAL
CLEF 2003	3	5	0	0	8
CLEF 2004	1	17	0	0	18
CLEF 2005	1	22	1	0	24
CLEF 2006	4	24	2	0	30
CLEF 2007	3	17	1	1	22

Table 6. Number of submitted runs

	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
CLEF 2007	37	23	14

Also the number of submitted runs has decreased sensibly, from a total of 77 registered last year to 22 (see The geographical distribution has anyway remained almost the same, recording a new entry of a group from Australia. No participants took part to any Bulgarian tasks. Table 6). A breakdown of participants and runs, according to language, is shown in Table 4 (Section 2.3). As in previous campaigns, a larger number of people chose to participate in the monolingual tasks, which once again demonstrated to be more approachable.

In the following subsections a more detailed analysis 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.

3.2 Dutch as Target

For the Dutch subtask of the CLEF 2007 QA task, three annotators generated 200 questions organized in 78 groups so that there were 16 groups with one question, 21 groups with two, 22 with three and 19 groups with four questions. Among the 200 questions 156 were factoids, 28 definitions and 16 list questions. In total, 41 questions had temporal restrictions. Table 7 and Annotators were asked to create questions with answers either in Dutch Wikipedia or in the Dutch newspaper corpus, as well as questions without known answers. Of 200 questions, 186 had answers in Wikipedia, and 14 in the newspaper corpus. Annotators did not create NIL questions.

Table 8 below show the distributions of topic types for groups and expected answer types for questions.

Table 7. Distribution of topic types

Topic type	Number of topics
OBJECT	29
PERSON	18
ORGANIZATION	12
LOCATION	10
EVENT	19

Annotators were asked to create questions with answers either in Dutch Wikipedia or in the Dutch newspaper corpus, as well as questions without known answers. Of 200 questions, 186 had answers in Wikipedia, and 14 in the newspaper corpus. Annotators did not create NIL questions.

Table 8. Distribution of expected answers for questions

Expected answer type	Number of questions
OTHER	45
PERSON	38
TIME	32
OBJECT	25
LOCATION	25
COUNT	14
ORGANIZATION	13
MEASURE	8

This year, two teams took part in the QA track with Dutch as the target language: the University of Amsterdam and the University of Groningen. The latter submitted both monolingual and cross-lingual (English to Dutch) runs. The 5 submitted runs were assessed independently by 3 Dutch native speakers in such a way that each question group was assessed by at least two assessors. In case of conflicting assessments, assessors were asked to discuss the judgements and come to an agreement.

Most of the occurred conflicts were due to difficulties in distinguishing between *inexact* and *correct* answers. Table 9 above shows the evaluation results for the five submitted runs (three monolingual and two cross-lingual). The table shows the number of Right, Wrong, ineXact and Unsupported answers, as well as the percentage of correctly answered Factoids, Temporally restricted questions, Definition and List questions.

The best monolingual run (gron072NLNL) achieved accuracy of 25.5%, which is slightly less than the best results in the 2006 edition of the QA task. The same tendency holds for the performance on factoid and definition questions.

One of the runs contained as many as 23 unsupported answers—this might indicate a bug in the system.

Table 9. Results for Dutch as target

Run	R #	W #	X #	% F #U	% T [156]	% D [41]	% L [28]	NIL [16]	#	% [0]	CWS	Overall accuracy
uams071qrz	15	160	1	23	9.0	4.9	3.6	0	0	0	0.02	7.54
gron071NLNL	49	136	11	4	24.4	19.5	35.7	6.3	20	0	0.06	24.5
gron072NLNL	51	135	10	4	25.6	19.5	35.7	6.3	20	0	0.07	25.5
gron071ENNL	26	159	8	7	10.3	14.6	32.1	6.3	20	0	0.02	13
gron072ENNL	27	161	7	5	10.9	14.6	32.1	6.3	16	0	0.02	13.5

3.3 English as Target

160 Factoids (in groups) were requested, together with 30 definitions and ten lists. The numbers of temporally restricted factoids and questions with NIL answers was at our discretion. In the end we submitted 161 factoids, 30 definitions and nine lists. In previous years we have been obliged to devise a considerable number of temporally restricted questions and this has proved very difficult to do with the majority of them being very contrived and artificial. For this reason it was intended to set no such questions this year.

However, one reasonable one was spotted during the data entry process and so was flagged as such. Two others were also flagged accidentally during data entry. Unfortunately, therefore, the statistics cannot tell us anything about temporally restricted questions.

To achieve the goals set by the organizers it was necessary to find topics about which several questions could be asked and then to devise a set of questions from that topic. Each task was surprisingly hard, and an inevitable consequence was that the questions are much harder this year than in previous years. We had no wish to set especially difficult or convoluted questions, but unfortunately this arose as a side-effect of the new procedures.

In addition to the issue of question grouping, it was decided at a very late stage to use not only the two collections from last year (the LA Times and Glasgow Herald) but also the English Wikipedia. The latter is extremely large and greatly increases the task complexity for the participants in terms of both indexing and IR searching. In addition, some questions had to be heavily qualified in order to reduce the ambiguity introduced by alternative readings in the Wikipedia. Here is an example:

Q24: *What is the “KORG” on which Niky Orellana is a soccer commentator?*

The breakdown of the questions can be summarised as follows. There were 200 questions divided into 67 groups. In other words, there were 67 initial questions (33.50%) and 133 follow-on questions (66.50%) within the collection. Reference answers were established using the three collections. Of the 236 supporting snippets included in the corpus, 88 are from the LA Times (44.00%), 68 are from the Glasgow Herald (34.00%) and 44 are from the English Wikipedia (22.00%). Thus the majority of the reference answers were in the newspapers. However, as we shall see later, some systems found a much higher proportion of answers in the Wikipedia.

Table 10. Results for English as target

Run	R #	W #	X #	#U	% F [161]	% T [3]	% D [30]	% L [9]	NIL #	% [0]	CWS	K1	Overall accuracy
cind071fren	26	171	1	2	11.18	0.00	23.33	11.11	0	0.00	0.00	0.00	13.00
cind072fren	26	170	2	2	11.18	0.00	23.33	11.11	0	0.00	0.00	0.00	13.00
csui071inen	20	175	4	1	10.56	0.00	10.00	0.00	0	0.00	0.00	0.00	10.00
dfki071deen	14	178	6	2	4.35	0.00	23.33	0.00	0	0.00	0.00	0.00	7.00
dfki071esen	5	189	4	2	1.86	0.00	6.67	0.00	0	0.00	0.00	0.00	2:50

Five cross-lingual runs with English as target were submitted this year, as compared with thirteen for last year. Five groups participated in six source languages, Dutch, French, German, Indonesian, Romanian and Spanish. DFKI submitted runs for two source languages, German and Spanish, while all other groups worked in only one. Cindi Group and Macquarie University both submitted two runs for a language pair (French-English and Dutch-English respectively) but unfortunately there was no language for which more than one group submitted a run. This means that no direct comparisons can be made between QA systems this year, because the task being solved by each was different.

An XML format was used for the submission of runs this year, by contrast with previous years when fairly similar plain text formats were adopted. This meant that our evaluation tools were no longer usable. However, last year we also participated in the evaluation of the Question Answering using Wikipedia task (WiQA)³ organised by University of Amsterdam. For this they developed an excellent web-based tool which was subsequently adapted for this year's Dutch CLEF evaluations⁴. It allows multiple assessors to work independently, shows runs anonymised, allows all answers to a particular question to be judged at the same time (like the TREC software), and includes the supporting snippets for each submitted answer as well as the 'correct' (reference) answer. It also shows inter-assessor disagreement, and, once this has been eliminated, can produce the assessed runs in the correct XML format. Overall, this software worked perfectly for us and saved us a considerable amount of time.

All answers were double-judged⁵. Where assessors differed, the case was discussed between us and a decision taken. We measured the agreement level by two methods. For Agreement 1 we take agreement on each group of 8 answers to a question as a whole as either exactly the same for both assessors or not exactly the same. This is a very strict measure. There were disagreements for 30 questions out of the 200, i.e. 15%, which equates to an agreement level of 85%.

For Agreement Level 2 we taking each decision made on one of the eight answers to a question and count how many decisions were the same for both assessors and how many were not the same. There were 39 differences of decision and a total of 1600 decisions (200 questions by eight runs). This is 2.4%, which equates to an agreement level of 97.6%. This is the measure we used in previous years. Last year the agreement level was 89% and the previous year it was 93%. We conclude from these figures that the assessment of our CLEF runs is quite accurate and that double judging is sufficient.

Considering all question types together, the best performance is University of Wolverhampton with 28 R and 2 X, (14% strict or 15% lenient) closely followed by the CINDI Group at Concordia University with 26 R and 1 X (13% strict or 13.50% lenient). Note that these systems are working on different tasks (RO-EN and FR-EN respectively) as noted above, so the results are not directly comparable. The best performance last year for English targets was 25.26%. Nevertheless, considering the

³ <http://ilps.science.uva.nl/WiQA/>

⁴ We are extremely grateful to Martin de Rijke and Valentin Jijkoun for allowing us to use it and for setting it up in Amsterdam especially for us.

⁵ The first assessor was Richard Sutcliffe and the second was Udo Kruschwitz from University of Essex to whom we are indebted for his invaluable help.

extreme difficulty of the questions, this represents a remarkable achievement for these systems.

For Factoids alone, the best system was CINDI (FR-EN) at 11.18% followed by University of Indonesia (IN-EN) with 10.56%. For Definitions the best result was University of Wolverhampton (RO-EN) with 43.33% correct, followed equally by CINDI (FR-EN) and DFKI (DE-EN) both with 23.33%. It is interesting that this For Factoids alone, the best system was CINDI (FR-EN) at 11.18% followed by University of Indonesia (IN-EN) with 10.56%. For Definitions the best result was University of Wolverhampton (RO-EN) with 43.33% correct, followed equally by CINDI (FR-EN) and DFKI (DE-EN) both with 23.33%. It is interesting that this year the best Definition score is almost four times the best Factoid score, whereas last year they were nearly equal. One reason for this may be that the definitions either occurred first in a group of questions or on their own in a 'singleton' group. This was not specifically intended but seems to be a consequence of the relationship between Factoids and Definitions, namely that the latter are somehow epistemologically prior to the former⁶. In consequence, Definitions may be more simply phrased than Factoids and in particular may avoid co-reference in the vast majority of cases.

Nine list questions were set but only CINDI was able to answer any of them correctly (11.11% accuracy). (University of Indonesia was inexact on one list question.) Perhaps the problem here was recognising the list question in the first place – unlike at TREC they are not explicitly flagged.

Considering the runs collectively, only 119 correct answers were returned out of 1600 attempts (8 runs and 200 questions). 70 questions were answered correctly by at least one system and thus 130 were not answered by any system. Table 11 shows a breakdown of correct answers by the collection used by a system, and by the position of an answer in a particular question group. Taking the collections first, we can see that the systems used the Wikipedia much more than we might have expected. 22% of the reference answers came from the Wikipedia (see earlier) while here we see that the total figures (excluding DFKI) are 19/100 for Glasgow Herald (19%), 14/100 for LA Times (14%) and 67/100 for Wikipedia (67%). These figures suggest that many questions which were set relative to the newspapers were not answered from them. We will need to pay careful attention to this point before the next contest.

The last two columns in Table 11 show how good each system was at answering the first question in a group and the subsequent questions in a group – recall that there were 200 questions in 67 groups. As we can see, the number of subsequent questions answered correctly was less than the number of first questions. Across all the runs there were 68 correct answers to first questions (out of 67*8 attempts) and 51 correct answers to subsequent questions (out of 133*8 attempts). Thus the overall success rate on first questions was 12.69% and that on subsequent questions was 4.79%. This can be accounted for by the fact that subsequent questions are much more difficult because they use anaphoric references and also can involve knowing the answers to previous questions, as discussed earlier. The first column gives the number of correct answers returned by a system. The columns GH, LA and WI give the number of correct answers supported by snippets from the Glasgow Herald, LA Times and English

⁶ Perhaps it is just a consequence of setting too many undergraduate examination papers!

Table 11. Breakdown of the answers by collection and by the position of the question in a group

Run	All	GH	LA	WI	First	Subsq.
cind071fren	26	4	3	19	14	12
cind072fren	26	4	3	19	14	12
csui071inen	20	6	3	11	12	8
dfki071deen	14	-	-	-	9	5
dfki071esen	5	-	-	-	3	2
mqaf071nlen	0	0	0	0	0	0
mqaf072nlen	0	0	0	0	0	0
wolv071roen	28	5	5	18	16	12

Wikipedia respectively. The last two columns show the numbers of correct answers for initial questions in a group (First) and subsequent questions in a group (Subsq.). Information about DocIDs could not be extracted or inferred from the DFKI runs.

3.4 French as Target

This year two groups took part in evaluation tasks using French as target language: one French group: Synapse Développement; and one American group: Language Computer Corporation (LCC).

In total, only two runs have been returned by the participants: one monolingual run (FR-to-FR) from Synapse Développement and one bilingual run (EN-to-FR) from LCC.

It appears that the number of participants for the French task has clearly decreased this year, certainly due to the many changes that appeared in the 2007 Guidelines for the participants: adding to a large new answer source (the static version of Wikipedia, frozen in November 2006) and adding to a large number of topic-related questions. 200 answers were assessed for syn07frfr, and 194 for lcc0707enfr.

Figure 2 shows the best scores for systems using French as target in the last four CLEF QA campaigns.

Table 12. Results for French *as target*

Run	R #	W #	X #	U #	% F [161]	% T [3]	% D [30]	% L [9]	NIL #	% [0]	CWS	K1	Overall accuracy
syn07frfr	108	82	9	1	52.76	46.34	74.07	20	40	22.5	-	-	54 %
lcc07enfr	81	95	14	4	44.17	46.34	22.22	30	0	0	0.2223	-0.1235	41.75 %

The French test set was composed of 200 questions: 163 Factual (F), 27 Definition (D) and 10 closed List questions (L). Among these 200 questions, 41 were Temporally restricted questions (T).

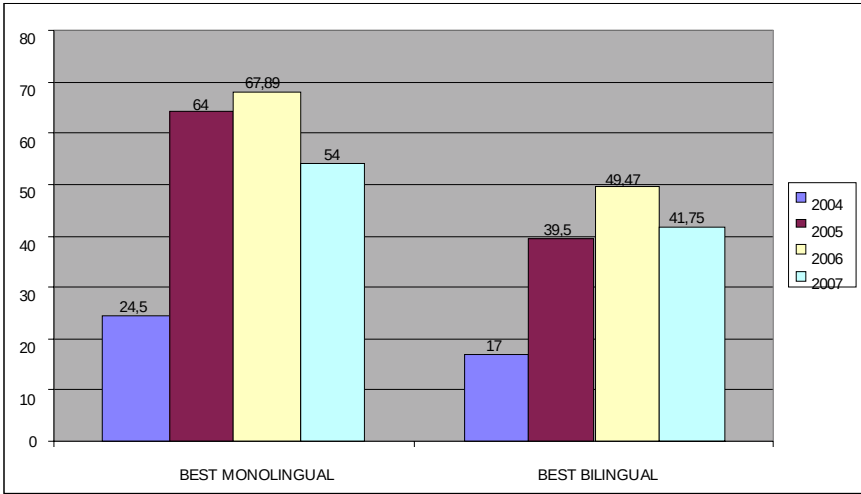


Fig. 2. Best scores for systems using French as target in QA@CLEF campaigns

The accuracy has been calculated over all the answers of F, D, T and L questions and also the Confidence Weighted Score (CWS) and the K1 measure.

For the monolingual task, the Synapse Développement' system returned 108 correct answers i.e. 54 % of correct answers (as opposed to 67,89 % last year).

For the bilingual task, the LCC's system returned 81 correct answers i.e. 41,75 % of correct answers (as opposed to 49,47 % for the best bilingual system last year).

We can observe that the two systems obtained different results according to the answer types. The monolingual system obtained better results for Definition questions (74,07 %) than for Factoid (52,76 %) and Temporally questions (46,34 %) whereas the bilingual system obtained better results for Temporally (46,34 %) and Factoid questions (44,17 %) than for Definition questions (22,22 %).

We can note that the bilingual system has not returned NIL answer, whereas the monolingual one returned 40 NIL answers (out of 9 expected NIL answers in the French test set). As there were only 9 NIL answers in the French test set and as the monolingual system returned 40 NIL answers, his final score is not very high (even if this system returned the 9 expected correct NIL answers).

In conclusion, despite the important changes in the Guidelines for the participants, the monolingual system obtained the best results of all the participants at CLEF@QA track this year (108 correct answers out of 200).

We can note that the American group (LCC) participated only for the second time in the Question Answering track using French in target and has already obtained good results that can let us imagine it will improve again in the future. In addition, we can still observe this year the increasing interest in Question Answering for the tasks using French as target language from the non-European research community due to the second participation of an American team.

3.5 German as Target

Two research groups submitted runs for evaluation in the track having German as target language: The German Research Center for Artificial Intelligence (DFKI) and the Fern Universität Hagen (FUHA).

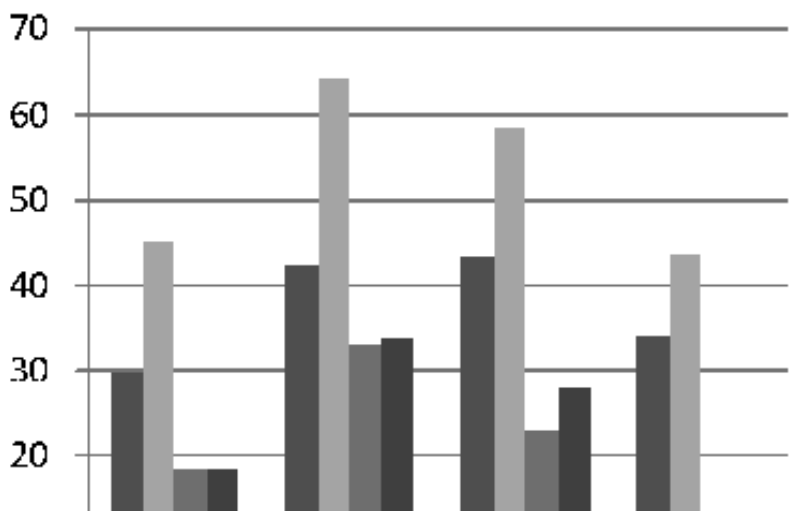


Fig. 3. Results evolution

Both provided system runs for the monolingual scenario and just DFKI submitted runs for the cross-language English-German and Portuguese-German scenario. Compared to the previous editions of the evaluation forum, this year a decrease in the accuracy of the best performing system and of an aggregated virtual system for both monolingual and cross-language tasks was registered, as seen in Figure 3.

The number of topics covered by the test set questions was of 116 distributed as it follows: 69 topics consisting of 1 question, each 19 topics of 2 and 3 related questions, and 9 topics of 4 related questions. The distribution of the topics over the document collections (CLEF vs. Wikipedia) is presented in Table 13.

Table 13. Results for German as target

Run	R #	W #	X #	U #	% F [164]	% T [27]	% D [28]	% L [8]	NIL #	% [0]	CWS	K1	Overall accuracy
dfki071dede _M	60	121	14	5	29.8	14.81	39.29	0	0	0	-	-	30
fuha071dede _M	48	146	4	2	24.39	18.52	28.57	0	0	0	0.086	-0.17	24
fuha072dede _M	30	164	4	2	17.07	14.81	7.14	0	0	0	0.048	-0.31	15
dfki071ende _C	37	144	18	1	17.68	14.81	25	12,5	0	0	-	-	18.5
dfki071ptde _C	10	180	10	0	3.66	7.41	14.29	0	0	0	-	-	5

According to Tables 14 and 15 the most frequent topic types were PERSON (40), OBJECT (33) and ORGANIZATION (23), with first two types more present for the news collection of documents (CLEF).

Table 14. Topic distribution over data collections

Topic Size	# Topics / CLEF	# Topics / WIKI	# Topics
1	53	16	69
2	4	15	19
3	4	15	19
4	7	2	9
Total	68	48	116

As regards the source of the answers, 101 questions from 68 topics asked for information out of the CLEF document collection and the rest of 99 from 48 topics for information from Wikipedia. Table 16 shows a breakdown of the test set questions by the expected answer type (EAType) for each collection of data.

Table 15. Topic type breakdown over CLEF collection

Topic Type	Topic Size				Total
	1	2	3	4	
PERSON	23	2	0	3	28
OBJECT	19	0	1	0	20
ORGANIZATION	8	1	1	2	12
LOCATION	1	1	1	0	3
EVENT	2	0	1	2	5
OTHER	0	0	0	0	0
					68

The system developed by DFKI relies on shallow NLP methods for both question and document processing and uses distance-based metrics and recall evidence for answer selection. The system developed by FUHA combines both shallow and deep NLP methods and uses semantic representations and an entailment engine for answer selection.

The details of systems’ results can be seen in Table 13. There were no NIL questions tested in this year’s evaluation. The results submitted by DFKI did not provide a normalized value for the confidence score of an answer and therefore both CWS and K1 values could not be computed.

A breakdown of results along self-contained questions, i.e. first ones in a topic with no reference to previous stated information – 116 in total, and linked questions, i.e.

Table 16. Topic type breakdown over Wikipedia collection

Topic Type	Topic Size				Total
	1	2	3	4	
PERSON	4	2	5	1	12
OBJECT	5	5	3	0	13
ORGANIZATION	3	3	5	0	11
LOCATION	2	1	1	1	5
EVENT	2	3	1	0	6
OTHER	0	1	0	0	1
					48

questions related to previous mentioned information or to the topic – 84 in total, shows a drop in the systems’ accuracy for the latter.

A thorough analysis of the questions unanswered by any of the participating systems revealed following common features of them:

- The answer’s context covers at least two sentences that might be adjacent (CLEF collection) or not (Wikipedia collection).
- The question and the answer’s context share semantic items, i.e. concepts, but not lexical items, i.e. words. Some examples of this phenomena are:
 - Ehe (marriage) vs verheiratet (married)
 - Geburtsname (birth name) vs bürgerlicher Name (civil name)
 - Band vs Popgruppe
 - Spielfilm von (motion picture by) vs verfilmt von (filmed by)
 - Beruf (profession) vs Rechtsanwalt (lawyer)
- The date asked for in question is not explicitly mentioned in the answer’s context, but assumed based on document’s publication date.

The assessment was conducted by two native German speakers with fair knowledge of information access systems. Table 17 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. Along the total figures for the disagreement, a breakdown at the

Table 17. Inter-assessor agreement/disagreement (breakdown)

Run	Number of questions	# Q-Disagreements						
		Total	F	D	L	X	U	W/R
dfki071dede _M	200	20	16	4	0	15	4	1
fuha071dede _M	200	13	10	3	0	7	3	3
fuha072dede _M	200	7	6	1	0	2	2	3
dfki071ende _C	200	13	7	5	1	12	1	0
dfki071ptde _C	200	8	3	5	0	8	0	0

question type level (Factoid, Definition, List) and at the assessment value level (ineX-act, 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.

3.6 Italian as Target

Only one group took part this year in the monolingual Italian task, i.e. FBK-irst, submitting only one run. The results are shown in Table 18.

Table 18. Results of the Italian monolingual task

Run	R	W	X	U	% F	% T	% D	% L	NIL	Correct	CWS	K1	Overall accuracy
	#	#	#	#	[161]	[3]	[30]	[9]	Returned				
irst071tit	23	160	4	13	15.17	12.5	2.63	0	14	3	0.017	0.043	11.55%

The Italian question set consisted of 147 factoid questions, 41 definition questions and 12 list questions. 38 questions contained a temporal restriction, and 11 had no answer in the Gold Standard. In the Gold Standard, 108 answers were retrieved from Wikipedia, the remains from the news collections (see Table 21). Results for Italian as target (answers to linked and unlinked questions). As Table 19 shows, the question set was almost perfectly balanced between questions were linked to a topic –which could contain co-references and needed to be considered as a group- and self-contained questions –which were similar to the queries proposed in the previous campaigns.

The submitted run was assessed by two judges; the inter-annotator agreement was 92,5%, meaning that the dataset contained a very low percentage of questionable cases.

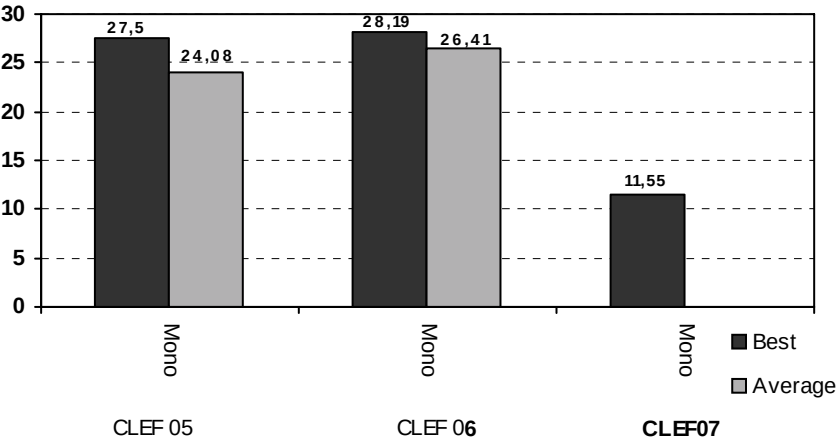


Fig. 4. Best and Average performance in the Monolingual and Bilingual tasks

As Figure 4 shows, the performance of the FBK-irst system was lower than those achieved in the previous campaigns: in 2006 the accuracy in the monolingual task was 22.87, almost twice as much as this year's score. Anyway, these results reflected the general trend also of the performances of the other systems largely due to the innovations introduced.

The system achieved low accuracy in all types of questions, performing somehow better in factoids questions. Definition questions, with 2.63% of accuracy and list questions, for which no correct answer was retrieved, proved to be particularly challenging.

Table 19. Results for Italian as target (answers to linked and unlinked questions)

	#	%	R	W	X	U
Question linked to a topic	108	54%	0	106	0	2
Self-contained questions	92	46 %	23	54	4	11
Total	200	100%	23	160	4	13

A relevant number of questions (about 6%) was judged unsupported, meaning that the correct answer was retrieved by the system, which did not provided enough context to support it.

Table 20. Results for Italian as target for NIL questions

	Precision (Overall)	Recall (Overall)
FBK-irst	0.21	0.27

Regarding the questions with no answers, the system returned the value NIL 14 times, compared to the 11 present in the Gold Standard. Therefore, as Table 20 shows, the overall precision about NIL questions was 0.21, with an overall recall of 0.27, which proves that NIL questions are still problematic.

It may be interesting to have a closer look at the results according to the new features introduced in this year's competition.

Table 21. Questions by source

Source	# Gold Standard	# FBK-irst	FBK-irst %	R	W	X	U
News	81	178	89%	19	142	4	13
Wikipedia	108	8	4%	1	7	0	0
Other (NIL)	11	14	7%	3	11	0	0
Total	200	200	100%	23	160	4	13

Meanwhile the answers in the Gold standard were almost equally retrieved from news collections and Wikipedia (see Table 21), the system found the answers mainly in the news collections, for a total of 178 out of 200, compared to the 8 responses extracted from Wikipedia. If we consider that in the Gold standard the answer retrieved from Wikipedia were 108, we could conclude that the system did not exploit this source properly. The reason for that should be probably investigated a bit longer. As for the precision of the answer with respect to the collections, it was 0.13 on Wikipedia and 0.11 on the news collection.

3.7 Portuguese as Target

Six research groups took part in tasks with Portuguese as target language, submitting eight runs: seven in the monolingual task, and one with English as source; unlike last year, no group presented Spanish as source. One new group (INESC) participated this year. The group of University of Évora (UE) returned this year, while the group from NILC, the sole Brazilian group to take part to date, was absent.

Again, Priberam presented the best result for the third year in a row; the group of the University of Évora wasn't however far behind. As last year, we added the classification X-, meaning incomplete, while keeping the classification X+ for answers with extra text or other kinds of inexactness. In Table 22 we present the overall results.

A direct comparison with last year's results is not fully possible, due to the existence of multiple questions to each topic. Therefore, in Question 94 was reclassified as NIL due to a spelling error, and question 135 because of the use of a rare meaning of a word. On the other hand, one system saw through that rare meaning, providing a correct answer; we decided to keep the question as NIL, considering correct both the system's answer and any NIL answer from other systems.

Table 23 we present the results both for first question of each topic (which we believe is more readily comparable to the results of previous years) and for the linked questions.

As it can be seen, apart from Priberam, the results over linked questions aren't much different from those over not-linked. On the whole, compared to last year [12], Priberam saw a slight drop on its results, Raposa (FEUP) a clear improvement from an admittedly low level, Esfinge (SINTEF) a clear drop, and LCC kept last year's levels. Senso (UE) shows a marked improvement since its last participation in 2005 [16].

Table 22. Results for Portuguese as target (all 200 questions)

Run	R	W	X+	X-	U	Overall	Accuracy	
	#	#	#	#	#	accuracy	Precision (%)	Recall (%)
diue071ptpt	84	103	1	11	1	42.0	11.7	92.3
esfi071ptpt	16	178	0	4	2	8.0	6.3	69.2
esfi072ptpt	12	184	0	2	2	6.0	6.1	84.6
feup071ptpt	40	158	1	1	0	20.0	8.3	84.6
ines071ptpt	22	171	1	4	2	11.0	7.3	69.2
ines072ptpt	26	168	0	4	2	13.0	7.2	84.6
prib071ptpt	101	88	5	5	1	50.5	27.8	46.2
lcc_071enpt	56	121	7	3	13	28.0	33.3	23.1

Question 94 was reclassified as NIL due to a spelling error, and question 135 because of the use of a rare meaning of a word. On the other hand, one system saw through that rare meaning, providing a correct answer; we decided to keep the question as NIL, considering correct both the system's answer and any NIL answer from other systems.

Table 23. Results for Portuguese as target (answers to linked and unlinked questions)

Run	First questions [149]						Linked questions [51]	
	R	W	X+	X-	U	Accuracy	R	Accuracy
	#	#	#	#	#	%	#	%
diue071ptpt	61	77	1	9	1	40.9	23	45.1
esfi071ptpt	11	132	0	4	2	7.4	5	9.8
esfi072ptpt	6	141	0	1	1	4.0	6	11.8
feup071ptpt	34	113	1	1	0	22.8	6	11.8
ines071ptpt	17	125	1	4	2	11.4	5	9.8
ines072ptpt	21	122	0	4	2	14.1	7	13.7
prib071ptpt	92	86	3	5	1	61.7	9	17.6
lcc_071enpt	44	48	7	3	9	29.5	12	23.5

The same system also found a correct answer to a question classified as NIL in the test set; that question was therefore reclassified as non-NIL. In the end, there were 13 NIL questions. Table 24 shows the results for each answer type of definition questions, while Table 25 shows the results for each answer type of factoid questions (including list questions). As it can be seen, four out of six systems perform clearly better when it comes to definitions than to factoids. This may well have been helped by the use of Wikipedia texts, where a large proportion of articles begin with a definition.

We included in both Table 24 and Table 25 a virtual run, called combination, in which one question is considered correct if at least one participating system found a valid answer. The objective of this combination run is to show the potential achievement when combining the capacities of all the participants. The combination run can be considered, somehow, state-of-the-art in monolingual Portuguese question answering. The system with best results, Priberam, answered correctly 72.7% the questions with at least one correct answer, not as dominating as last year; in all, 137 questions had at least one correct answer among the monolingual runs (67% of first questions and 47% of linked questions); 75 questions were answered by more than one system, and only four (all NIL) by all monolingual runs.

Despite being a bilingual run, LCC answered correctly to 14 questions not answered by any of the monolingual systems.

Analysing those questions which no system managed to answer, and comparing them with the test set extract chosen by the organization, it seems that the most important cause are the non-handling of anaphora - both in the questions (while only 20% of the first questions of each topic found no correct answer, that number rises to 37% of the subsequent questions), and of the collection text itself (e.g., questions 170 and 172).

Table 24. Results of the assessment of the monolingual Portuguese runs: definitions

Run	obj	org	oth	per	TOT %	
	5	22	28	24		
diue071ptpt	6	4	5	4	19	63%
esfi071ptpt	1	0	0	0	1	3%
esfi072ptpt	1	0	0	0	1	3%
feup071ptpt	3	2	4	7	16	53%
ines071ptpt	4	4	6	0	14	47%
ines072ptpt	5	5	6	2	18	60%
prib071ptpt	6	4	6	7	23	77%
combination	6	5	8	9	27	87%
lcc_071enpt	2	3	2	1	8	27%

Secondary issues yet to be fully tackled are the use of the document date when not mentioned in the document (e.g., questions 71, 128, 168 and 194), of using the common root of words to find an answer (questions 11 and 196), of validating dates when intervals are used in the text (questions 55 and 140) and of finding Portuguese equivalents for Brazilian terms or vice-versa (questions 21 and 98).

Table 25. Results of the assessment of the monolingual Portuguese runs: factoids, including lists

Run	cou	loc	Mea		obj	org	oth	per	tim	TOT %	
	21	34	16	5	22	28	24	20			
diue071ptpt	11	17	4	3	6	8	7	9	65	38%	
esfi071ptpt	3	3	0	0	1	0	1	7	15	9%	
esfi072ptpt	2	4	0	0	1	0	2	2	11	6%	
feup071ptpt	4	8	0	0	3	1	3	5	24	14%	
ines071ptpt	1	3	0	0	0	0	2	2	8	5%	
ines072ptpt	2	4	0	0	0	0	2	2	10	6%	
prib071ptpt	9	15	10	1	11	14	8	10	78	46%	
combination	16	24	12	3	12	17	12	13	109	64%	
lcc_071enpt	7	11	6	1	3	10	4	6	48	28%	

In Table 27 presents the results of the 20 temporally restricted questions. As in previous years, the effectiveness of the systems to answer those questions is visibly lower than for non-TRQ questions (and indeed most systems only answered correctly question 160, which is a NIL TRQ).

Table 26 we present some values concerning answer and snippet size. Table 27 presents the results of the 20 temporally restricted questions. As in previous years, the effectiveness of the systems to answer those questions is visibly lower than for non-TRQ questions (and indeed most systems only answered correctly question 160, which is a NIL TRQ).

Table 26. Average size of answers (values in number of words)

Run name	Non-NIL	Average	Average answer	Average	Average snippet
	Answers	answer	size (R only)	snippet	size (R only)
	#	size		size	
diue071ptpt	89	2.8	2.9	25.0	24.3
esfi071ptpt	57	2.4	2.8	56.3	29.3
esfi072ptpt	19	2.4	2.8	59.7	29.1
feup071ptpt	56	2.7	3.3	59.8	32.9
ines071ptpt	49	3.7	4.8	60.7	33.6
ines072ptpt	47	3.8	5.3	61.7	34.2
prib071ptpt	182	3.5	4.4	49.6	32.4
lcc_071enpt	191	3.4	4.2	45.2	32.7

A total of twelve questions were defined as list questions; unlike last year, all these questions were closed list factoids, with two to twelve answers each⁷.

The results were, in general, weak, with UE and LCC getting two correct answers, Priberam five, and all other system zero. There was a single case of incomplete answer (i.e., answering some elements of the list only), but it was judged W since, besides incomplete, it was also unsupported.

Table 27. Accuracy of temporally restricted questions

Run name	Correct answers	T.R.Q.	Non-T.R.Q.	Total
		correctness	correctness	correctness
	#	%	%	%
diue071ptpt	4	20.0	44.4	42.0
esfi071ptpt	1	5.0	8.3	8.0
esfi072ptpt	1	5.0	6.1	6.0
feup071ptpt	1	5.0	21.7	20.0
ines071ptpt	1	5.0	11.7	11.0
ines072ptpt	1	5.0	15.0	14.0
prib071ptpt	8	40.0	51.7	28.0
lcc_071enpt	6	30.0	27.8	50.5

Table 28 presents the distribution of questions by source during their selection, while Table 29 presents the distribution of sources used by the different runs and their correctness.

As it can be seen, the systems found the answers to half of the questions originally selected from newswire in Wikipedia (27 out of 55); conversely, only 5% of questions selected from Wikipedia received a correct answer from newspaper sources.

⁷ There were some open list questions as well, but they were classified and evaluated as ordinary factoids.

Table 28. Questions by source

Source	# during selection	# including valid answers
Wikipedia	132	159
News	55	62
NIL	13	13

Table 29. Answers by source and their correctness

Run	News	%	Wikipedia	%	NIL	%
	#	correct	#	Correct	#	correct
diue071	10	80%	79	81%	111	11%
esfi071	53	9%	4	50%	143	6%
esfi072	18	6%	1	0%	181	6%
feup071	17	71%	39	44%	144	8%
ines071	30	13%	23	39%	147	6%
ines072	28	14%	23	57%	149	7%
prib071	41	63%	141	50%	18	28%
lcc_071	17	24%	174	30%	9	0%

3.8 Romanian as Target

The creation of the questions was realized at the Faculty of Computer Science, Al.I. Cuza University of Iasi. The group⁸ was very well instructed with respect to this task, using the Guidelines for Question Generation and based on a good feedback received from the organizers at IRST⁹. The final 200 created questions are distributed according to Table 30 where for each type of question and expected answer we indicate also the temporally restricted questions out of the total number of questions. For Romanian, as source and target language we used only the collection of Wikipedia articles, hence the answers of 100% of the questions are in Wikipedia (without counting the NIL questions).

Table 30. Question and answer types distribution in Romanian (in brackets the number of temporally restricted questions)

Q type / expected A type	PERS	TIME	LOCAT	ORG	MEA	COU	OBJ	OTH	TOTAL
FACTOID	22 (14)	17	21 (4)	19 (8)	17	20 (7)	16 (6)	21 (6)	153 (45)
DEFINITION	9	-	-	5	-	-	6 (1)	10 (1)	30 (2)
LIST	5 (1)	-	2	-	-	-	1	2	10 (1)
NIL	3 (1)	-	-	-	-	1 (1)	2 (1)	1	7 (3)

⁸ Three Computational Linguistics Master students: Anca Onofraş, Ana-Maria Rusu, Cristina Despa, supervised and working in collaboration with the two organizers.

⁹ Without the help received from Danilo Giampiccolo and Pamela Forner, we wouldn't have solved all our problems.

We decided to include NIL questions, even though they seem somehow unnatural; the way we created them was not by including questions about facts impossible from a human perception. The Romanian NIL questions have answers in the English online Wikipedia, but not in the frozen Romanian Wikipedia articles.

This year in QA@CLEF one novelty were the questions related under the same topic: the organizers had to choose a certain number of topics and to create up to four questions related under one same topic. Using also the classification available within the question generation upload interface¹⁰, the percentage of topic-linked questions is illustrated in Table 31. This table shows that 129 questions were grouped under 51 topics, hence 64.5% out of the total 200 questions were linked in under topics with more than one question.

Most difficulties in this task were raised by deciding on the supporting snippets, especially for questions belonging to the same topic. We found unnatural to include answers through “copy-paste” from the text, because this way the answer was grammatically incorrect in some situations.

Table 31. Percentage of topic-linked questions

# of questions / Topic type	PERSON	LOCATION	ORGANIZ.	EVENT	OBJECT	OTHER	TOTAL
4 Qs	4		1			1	6
3 Qs	6	1	1		4	3	15
2 Qs	11	5	4	2	3	5	30
1 Q	14	7	15	3	11	21	71
TOTAL	35	13	21	5	18	30	122

For the LIST question we prepared also some questions with the answer to be found in various sections of an article or even in various articles. The situation is plausible from the point of view of an user asking for automatic answers.

We illustrate only the first type of LIST question with the following example: for the question *Name the main laws initiated by Cuza*. (RO: *Numiți prinipalele legi inițiate de Cuza.*), the answer should be extracted from various sentences in the same article¹¹. We show (underlined) only the sentences from where the answer should be extracted: [...] se întocmește un Proiect de lege organică pentru instrucția publică în Principatele Unite, [...] Noul guvern prezintă Adunării și realizează proiectul legii privind secularizarea averilor mănăstirești, lege prin care s-a dat o lovitură puternică feudalismului. De asemenea, se supune poporului, spre aprobare prin plebiscit, o nouă contribuție, o nouă lege electorală. [...] În acest an se decretează Legea Rurală, prin care se desființează iobăgia. Reforma agrară din 1864, a cărei aplicare s-a încheiat în linii mari în 1865, a satisfăcut în parte setea de pământ a țăranilor, [...]. The English version¹² of the same Wikipedia article includes even more laws: *His first measure*

¹⁰ http://www.celct.it/Question_generation_interface/question_generation_interface.html

¹¹ [/ro/a/l/e/Alexandru_Ioan_Cuza_9c42.html](http://ro/a/l/e/Alexandru_Ioan_Cuza_9c42.html)

¹² http://en.wikipedia.org/wiki/Alexandru_Ioan_Cuza

addressed a need for increasing the land resources and revenues available to the state, by "secularizing" (confiscating) monastic assets (1863). [...] The land reform, liberating peasants from the last *corvées*, freeing their movements and redistributing some land (1864), was less successful. [...] His plan to establish universal manhood suffrage, together with the power of the Domnitor to rule by decree, passed by a vote of 682,621 to 1,307. He consequently governed the country under the provisions of Statutul de dezvoltător al Convenției de la Paris ("Statute expanding the Paris Convention"), an organic law adopted on July 15, 1864. With his new plenary powers, Cuza then promulgated the Agrarian Law of 1863. [...] Cuza's reforms also included the adoption of the Criminal Code and the Civil Code based on the Napoleonic code (1864), a Law on Education, establishing tuition-free, compulsory public education for primary schools. The examples show that the Romanian version includes 5 answers whereas the English one has 9 laws to be included in a list answer.

This year two Romanian groups took part in the monolingual task with Romanian as a target language: the Faculty of Computer Science from the Al. I. Cuza University of Iasi (UAIC), and the Research Institute for Artificial Intelligence from the Romanian Academy (RACAI), Bucharest. Three runs were submitted – one by the first group and two by the second group [14], with the differences between them due to the way they treated the question-processing and the answer-extraction.

The RACAI systems are based on the parse tree of the candidate sentence and are using different heuristics to match keywords from the questions with those of the sentence; they use the same corpus processing tool, TTL [7] - for tokenization, POS-tagging, lemmatization, NE recognition and chunking, LexPar [8] - for link analysis, the same text search engine (based on Lucene¹³) and different question analysis and answer extraction modules.

The UAIC system follows the traditional QA systems architecture: a corpus pre-processing module, a question analyser (including an anaphora resolution (AR) module, to handle topic-related questions), a module dedicated to index creation and Information Retrieval (based on the same Lucene), and an answer extractor. Next to the AR module, another novelty of the UAIC system is the use of a Textual Entailment module for the answer extraction.

The 2007 general results are presented in Tables 32, 33 and 34. The statistics includes a system, named *combined (0)*, obtained through the combination of the 3 participating RO-RO systems. This "ideal" system permits to calculate the percentage of the questions (and their type), answered by at least one of the three systems.

Table 32. Results in the monolingual task. Romanian as target language (I).

Run	R	W	X	U	Overall accuracy	NIL returned	NIL correct
combined (0)	81	91	37	1	40.5	7	7
outputRoRo (1)	24	171	4	1	12	100	5
ICIA071RORO (2)	60	105	34	1	30	54	7
ICIA072RORO (3)	60	101	39	0	30	54	7

¹³ <http://lucene.apache.org/>

All three systems crashed on the LIST questions. The two RACAI systems did not include rules to handle this type of question [14], whereas the UAIC system had a simple rule (if the question focus is a plural noun, then the question type is LIST).

Table 33. Results in the monolingual task. Romanian as target language (II).

Run	Factoid Questions				List Questions				Definition Questions						
	R	W	U	X	ACC	R	W	U	X	ACC	R	W	U	X	ACC
(0)	52	76	1	84	33.98	0	10	0	0	0	22	5	0	3	73.33
(1)	24	131	1	2	15	0	10	0	0	0	0	30	0	0	0
(2)	38	90	1	31	23.75	0	10	0	0	0	22	5	0	3	73.33
(3)	38	86	0	36	23.75	0	10	0	0	0	22	5	0	3	73.33

The NIL questions are hard to classify, starting from the question-classifier (the classifier should “know” that the QA system has no possibility, no knowledge to find the answer). It would be better to have a clear separation between the NIL answers due to impossibility to find answer and the NIL answers classified as such by the system. The performance of all the three systems with respect to the NIL questions is as high as indicated in Table 34 because the systems treated the questions non-classifiable in any of the other types (F, D or L) as NIL.

Table 34. Results in the monolingual task. Romanian as target language (III).

Run	Temporally Restricted					NIL				
	R	W	U	X	ACC	R	W	U	X	ACC
(0)	19	24	0	8	37.25	7	0	0	0	100
(1)	11	39	0	1	21.57	5	95	0	0	5
(2)	10	31	0	10	19.61	7	47	0	0	12.96
(3)	10	31	0	10	19.61	7	47	0	0	12.96

For the DEFINITION questions the UAIC system considered them as such if the expected answer is of type D, whereas the answer classifier is based on patterns, specific for each type of answer. The RACAI systems are using dedicated rules for the D questions, hence the performance is understandable. The D answers judged as X or W are due to too long answers, too short snippets or to snippets that are shortened as such as they do not include the Right answer. For example for the question *Ce este Selena?* (EN: *What is Selene?*), the answer returned by the RACAI systems was: *o actriță și cântăreață americană , născută pe 24 iulie 1969 , în cartierul Bronx din New York* (EN: *an American actress and singer, born on July 24, 1969 in Bronx, New York*). The answer is considered “good enough” [14], but it was judged as wrong because it indicates the actress who played the role of Selena in the homonymous movie. The correct answer is *satelitul natural al Pamântului* (EN: *the natural satellite of the Earth*). The answer returned by the systems could reply to the D question “*What is Jennifer López?*”, according to the sentence in the wikipedia article and the

provided snippet *Jennifer López este o actriță și cântăreață americană, născută pe 24 iulie 1969, în cartierul Bronx din New York* (EN: *Jennifer López is an American actress...*). The focus of the question (also the topic of the group of questions) is *Selena*, which anyway is not a defined entity in the text *Dar succesul a fost de partea ei abia în anul 1997, când a jucat rolul binecunoscutei și regretatei Selena, în filmul cu același nume*. (EN: *But her success came only in 1997 when she played the role of Selena, the famous and regretted person in the homonymous movie*). The topic of *Selena* proved to be for the RACAI systems a very good example of 3 topic-related questions for which the systems returned Right answers for the 2nd and 3rd questions, even though the first one had a wrong answer. The same situation appeared in many other topic-related questions answered by the RACAI systems, as we will show below. This proves that the strategy employed¹⁴ (adding to the query generated for a new question the query of the first question of the group, namely the topic of the question and the focus of the 10 first answers returned to the previous question of the group) is a good one.

The topic-related questions were handled by UAIC through a dedicated AR module able to work by identifying the antecedents of anaphors that refer to a previous question answer or focus or by expanding the keywords lists of the questions in a same group with the keywords of the first question in that group. This strategy allowed identifying an answer as X in one case, as U in another one and as R in 9 cases of topic-related questions (the first one in the group is excluded). In 5 of these cases the R answer is NIL, hence the AR strategy was not used. For the other 4 cases the answer was R because the strategy worked and the system has specially developed rules for the MEASUREMENT answers (one case), for the temporally restricted questions (one case) or the question contains many keywords. Therefore the UAIC percentage of R answers for linked questions is 6.97%. The RACAI strategy for linked questions conducted to 20 R answers (15.5%), 15 of type X (11.62%) and 1 – U for the 2nd, 3rd or even the 4th question in a topic-related group. Six of the 20 R answers are for NIL questions, hence no strategy was used but only for the other 14 questions. One very nice such example has the topic *International Monetary Fund*, where the first question (*Which organization was formed in 1945 with the purpose of promoting a healthy global economy?*) included the topic only in the expected answer, not found by the RACAI systems. But for the second question (*How many members does it have?*) the answer is right (184).

The RACAI answers were judged as X or U, and not only for the topic-related questions, mainly due to answers that are too long, snippets shortened as such as they do not contain the answer (in fact in some situations the answer is the only one missing from the snippet) or because there are cases where the answer and the snippet has no connections (the answer extraction module). The UAIC answers of type X and/or U were judged as such mainly because the snippets are too long and they do not contain full clauses, but segments of clauses or sentences, unexpectedly stopped.

Due to time restrictions, all three runs were judged by only one assessor at the Faculty of Computer Science in Iasi, so an inter-annotator agreement was not possible. Based on the Guidelines, all three systems were judged in parallel. The same

¹⁴ We thank to prof. Dan Tufis for clarifying the methodology.

evaluation criteria, especially with respect to the U and X answers, were used. The analyses described above are based on a thorough manual introspection.

3.9 Spanish as Target

The participation at the Spanish as Target subtask has decreased from 9 groups in 2006 to 5 groups this year. All the runs were monolingual. We think that the changes in the task (linked questions and Wikipedia) led to a lower participation and worse overall results because systems could not be tuned on time.

Table 35 shows the summary of systems results with the number of Right (R), Wrong (W), Inexact (X) and Unsupported (U) answers. The table shows 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.

All the runs were assessed by two assessors. Only a 1.5% of the judgements were different and the resulting kappa value was 0,966, which corresponding to “almost perfect” assessment [10].

Table 35. Results for Spanish as target

Run	R #	W #	X #	U #	% F [115]	% T [43]	% D [32]	% L [10]	NIL #	F [8]	CWS	K1	Overall accuracy
Priberam	89	87	3	21	47,82	23,25	68,75	20	3	0,29	-	-	44,5
Inaoe	69	118	7	6	28,69	18,60	87,50	-	3	0,12	0,175	-0,287	34,5
Miracle	30	158	4	8	20	13,95	3,12	-	1	0,07	0,022	-0,452	15
UPV	23	166	5	6	13,08	9,30	12,5	-	1	0,03	0,015	-0,224	11,5
TALP	14	183	1	2	6,08	2,32	18,65	-	3	0,07	0,007	-0.34	7

Table Table 36 shows some evidence on the effect of Wikipedia in the performance. When the answer appears only in Wikipedia the accuracy is reduced in more than 35% in all the cases. Regarding NIL questions, The correlation coefficient r between the self-score and the correctness of the answers (shown in Table 39), has been similar to the obtained last year, being not good enough yet, and explaining the low results in CWS and K1 [6] measures.

Table 37 shows the harmonic mean (F) of precision and recall for self-contained, linked and all questions.

The best performing system has decreased their overall performance with respect to the last edition (see Table 38). in NIL questions. However, the performance considering only self-contained questions is closer to the one obtained last year.

The correlation coefficient r between the self-score and the correctness of the answers (shown in Table 39), has been similar to the obtained last year, being not good enough yet, and explaining the low results in CWS and K1 [6] measures.

Table 36. Results for self-contained and linked questions, compared with overall accuracy

Run	% Accuracy over Self-contained questions	% Accuracy over Linked questions	% Overall Accuracy
	[170]	[30]	[200]
Priberam	49,41	16,66	44,5
Inaoe	37,64	16,66	34,5
Miracle	15,29	13,33	15
UPV	12,94	3,33	11,5
TALP	7,05	6,66	7

Table 37. Results for Spanish as target for NIL questions

	F-measure (Self-contained)	F-measure (Overall)	Precision (Overall)	Recall (Overall)
Priberam	0.4	0.29	0.23	0.38
Inaoe	0.13	0.12	0.07	0.38
Miracle	0.07	0.07	0.05	0.13
UPV	0.04	0.03	0.02	0.13
TALP	0.06	0.07	0.04	0.38

Since a supporting snippet is requested in order to assess the correctness of the answer, we have evaluated the systems capability to extract the answer when the snippet contains it.

Table 38. Evolution of best results for NIL questions

Year	F-measure
2003	0,25
2004	0,30
2005	0,38
2006	0,46
2007	0,29

The first column of Table 39 shows the percentage of cases where the correct answer was present in the snippet and 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 process. As shown in the table, the best systems are also better in the task of answer extraction, whereas the rest of systems still have a lot of room for improvement.

Table 39. Answer extraction and correlation coefficient (r) for Spanish as target

Run	%Answer Extraction	R
Priberam	93,68	-
INAOE	75	0,1170
Miracle	49,18	0,237
UPV	54,76	-0,1003
TALP	53,84	0,134

4 Final Analysis

This year the task was changed considerably and this affected the general level of results and also the level of participation in the task. The grouped questions could be regarded as more realistic and more searching but in consequence they were much more difficult. The policy of not declaring the question type means that if this is deduced incorrectly then the answer is bound to be wrong. Moreover, the policy of not even declaring the topic of a question group, but leaving it implicit (usually within the first question) means that if a system infers the topic wrongly, then all questions in the group will be answered wrongly. Neither of these strike us as particularly ‘realistic’. In a real dialogue, if a question is answered inappropriately we do not dismiss all subsequent answers from that person, we simply re-phrase the question instead. The level of ambiguity concerning question type in a real dialogue is not fixed at some arbitrary value but varies according to many factors which the questioner estimates. In CLEF we are not modelling this process at all accurately and this affects the validity of our results. In addition, co-reference has now entered CLEF. This is interesting and useful but it might be preferable if we could separate the effect of co-reference resolution from other factors in analysing results. This could be done by marking up the co-references in the question corpus and allowing participants to use this information under certain circumstances. Finally, we have for the first time used the Wikipedia as a source of questions. For English targets there were few questions intended to be answered from it, but in practice many of the returned answers were supported by Wikipedia snippets. We could interpret this in different ways. On the one hand, we could argue that it shows how good Wikipedia is at answering simple questions, from which it follows that the newspaper corpora could be discarded. An alternative point of view, however, could be that it is valuable to be able to extract *additional* knowledge from newspapers and that therefore the Wikipedia could be excluded from certain tasks. This is a point which needs further discussion.

From the analyses accomplished by the organizing groups for German, Portuguese and Spanish, an overall decrease in the accuracy reached by the systems when treating linked questions can be observed. This fact evidences that topic resolution seems to be a weak point for QA systems. In the present edition topic-related questions were proposed for the first time and the participants did not have much time to tune their systems. As a consequence, they could not manage as well as in previous editions. There exist evidences that the most important cause is the non-handling of anaphora,

as referred the team in charge of Portuguese after an analysis of the data related to their language. From the questions which no system managed to answer for Portuguese as target language, only 20% of the first questions of each topic found no correct answer. But, that number rises to 37% of the subsequent questions.

Another source of difficulties, as referred by some participants, is the inclusion of Wikipedia as document corpus. These participants argue that the overall decrease in the accuracy reached by their systems comes from several problems when consulting Wikipedia. In all cases, these problems are a consequence of the impossibility of tuning the systems to the new requirements of the task in the time available. As instance, Synapse [11] could not adapt its system to a pattern extraction from Wikipedia as accurate as the one implemented for the news corpus. University of Hagen [5] found problems when treating article names, which led to an inconsistent concept index that rendered many Wikipedia articles inaccessible for its system.

In addition, the drop of the number of participating teams caused that, for certain pairs of source and target languages, one team tackled the subtask. Therefore, a comparison between systems working under the same circumstances cannot be accomplished. It impedes one of the major goals of campaigns such the QA@CLEF: the systems comparison in order to determine better approaches.

5 Future Work

After 5 years experiencing with QA issues, a lot of resources and know-how is accumulated nowadays. But systems do not show a brilliant overall performance, even those that participate edition by edition. The systems evidenced that they could not manage suitably the challenges proposed in the present edition while improving their performance when tackling issues already treated in previous campaigns. Given this situation, perhaps is time for no more innovation in question and answer types but for revising, little by little, every aspect considered until now in the past campaigns, in order to stimulate the improvement of the systems in a few skills every year. For this, without forgetting that nowadays sufficient evaluation resources from the previous years are available, in following campaigns a new focusing could be given to the task, as instance:

- Component evaluation, i.e., question classification, topic resolution, passage retrieval, answer extraction or answer validation (the latter already developed in the AVE).
- Join some target languages into a single multilingual target collection. Portuguese and Spanish are good candidates since they are closed languages and have many participants.
- Evaluation of an only question type every year.

In addition, being the development of high-performance QA systems a desirable goal, not only an accurate definition of every task should be accomplished but a more in-depth analysis of the participant systems, in order to determine relations between implementations and results.

Acknowledgements. A special thank to Bernardo Magnini (FBK-irst, Trento, Italy), who has given his precious advise and valuable support at many levels for the preparation and realization of the QA track at CLEF 2007.

Jesús Herrera has been partially supported by the by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project).

Anselmo Peñas has been partially supported by the Spanish Ministry of Science and Technology within the Text-Mess-INES project (TIN2006-15265-C06-02).

Paulo Rocha was supported by the Linguatca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC.

References

1. QA@CLEF Website, <http://clef-qa.itc.it/>
2. AVE Website, <http://nlp.uned.es/QA/ave/>
3. QAST Website, <http://www.lsi.upc.edu/~qast/>
4. QA@CLEF 2007 Organizing Committee. Guidelines (2007), http://clefqa.itc.it/2007/download/QA@CLEF07_Guidelines-for-Participants.pdf
5. Hartrumpf, S., Glöckner, I., Leveling, J.: University of Hagen at QA@CLEF 2007: Coreference Resolution for Questions and Answer Merging. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
6. Herrera, J., Peñas, A., Verdejo, F.: Question Answering Pilot Task at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 581–590. Springer, Heidelberg (2005)
7. Ion, R.: Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis, Romanian Academy, Bucharest (2007)
8. Ion, R., Mititelu, V.B.: Constrained Lexical Attraction Models. In: Nineteenth International Florida Artificial Intelligence Research Society Conference, pp. 297–302. AAAI Press, Menlo Park (2006)
9. Jijkoun, V., de Rijke, M.: Overview of the WiQA Task at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 265–274. Springer, Heidelberg (2007)
10. Landis, J.R., Koch, G.G.: The measurements of observer agreement for categorical data. *Biometrics* 33, 159–174 (1997)
11. Laurent, D., Séguéla, P., Nêgre, S.: Cross Lingual Question Answering using QRISTAL for CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
12. 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 223–256. Springer, Heidelberg (2007)
13. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
14. Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A.: RACAI's Question Answering System at QA@CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)

15. Turmo, J., Comas, P., Ayache, C., Mostefa, D., Rosset, S., Lamel, L.: Overview of QAST 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
16. 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)
17. Voorhees, E.: Overview of the TREC 2002 Question Answering Track. In: The Eleventh Text REtrieval Conference (TREC 2002), National Institute of Standards and Technology, USA, NIST Special Publication 500-251 (2002)

Overview of the Answer Validation Exercise 2007

Anselmo Peñas, Álvaro Rodrigo, and Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{anselmo, alvarory, felisa}@lsi.uned.es

Abstract. The Answer Validation Exercise at the Cross Language Evaluation Forum is aimed at developing systems able to decide whether the answer of a Question Answering system is correct or not. We present here the exercise description, the changes in the evaluation methodology with respect to the first edition, and the results of this second edition (AVE 2007). The changes in the evaluation methodology had two objectives: the first one was to quantify the gain in performance when more sophisticated validation modules are introduced in QA systems. The second objective was to bring systems based on Textual Entailment to the Automatic Hypothesis Generation problem which is not part itself of the Recognising Textual Entailment (RTE) task but a need of the Answer Validation setting. 9 groups have participated with 16 runs in 4 different languages. Compared with the QA systems, the results show an evidence of the potential gain that more sophisticated AV modules introduce in the task of QA.

1 Introduction

The first Answer Validation Exercise (AVE 2006) [7] was activated last year in order to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by QA systems. In some sense, systems must emulate human assessment of QA responses and decide whether an answer is correct or not according to a given text. This automatic Answer Validation is expected to be useful for improving QA systems performance [5]. However, the evaluation methodology in AVE 2006 did not permit to quantify this improvement and thus, the exercise has been modified in AVE 2007.

Figure 1 shows the relationship between the QA main track and the Answer Validation Exercise. The main track provides the questions made by the organization and the responses given by the participant systems once they are judged by humans.

Another difference in the exercise with respect to the AVE 2006 is the input to the participant systems. Last year we promoted an architecture based on Textual Entailment trying to bring research groups working on machine learning to Question Answering. Thus, we provided the hypothesis already built from the questions and answers [6] (see Figure 2). Then, the exercise was similar to the RTE Challenges [1] [2] [3], where systems must decide if there is entailment or not between the supporting text and the hypothesis.

In this edition, on the contrary, we left open the problem of Automatic Hypothesis Generation for those systems based on Textual Entailment. In this way, the task is more realistic and close to the Answer Validation problem, where systems receive a triplet (Question, Answer, Supporting text) instead a pair (Hypothesis, Text) (see Figure 2).

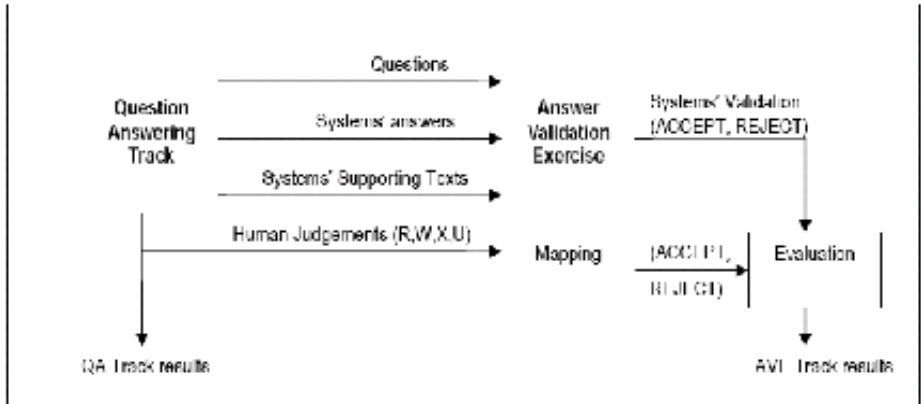


Fig. 1. Relationship between the QA Track and the AV Exercise

Section 2 describes the exercise in more detail. The development and testing collections are described in Section 3. Section 4 discusses the evaluation measures. Section 5 offers the results obtained by the participants and finally Section 6 present some conclusions and future work.

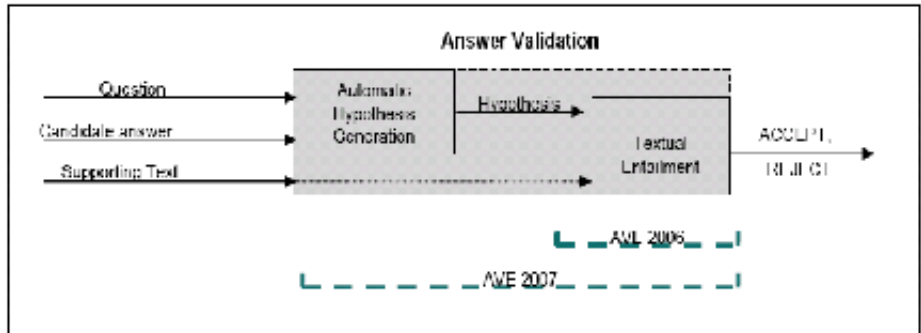


Fig. 2. From an Answer Validation architecture based on Textual Entailment in AVE 2006 to the complete Answer Validation systems evaluation in AVE 2007

2 Exercise Description

In this edition, participant systems received a set of triplets (Question, Answer, Supporting Text) and they must return a value for each triplet rejecting or accepting it. More in detail, the input format was a set of pairs (Answer, Supporting Text) grouped by Question (see Figure 3). Systems must consider the Question and validate each of the (Answer, Supporting Text) pairs. The number of answers to be validated per question depended on the number of participant systems at the Question Answering main track.

```

<q id="116" lang="EN">
  <q_str>What is Zanussi?</q_str>
  <a id="116_1" value="">
    <a_str>was an Italian producer of home appli-
      ances</a_str>
    <t_str doc="Zanussi">Zanussi For the Polish film di-
      rector, see Krzysztof Zanussi. For the hot-air bal-
      loon, see Zanussi (balloon). Zanussi was an Italian
      producer of home appliances that in 1984 was
      bought</t_str>
  </a>
  <a id="116_2" value="">
    <a_str>who had also been in Cassibile since August
      31</a_str>
    <t_str doc="en/p29/2998260.xml">Only after the sign-
      ing had taken place was Giuseppe Castellano informed
      of the additional clauses that had been presented by
      general Ronald Campbell to another Italian general,
      Zanussi, who had also been in Cassibile since August
      31.</t_str>
  </a>
  <a id="116_4" value="">
    <a_str>3</a_str>
    <t_str doc="1618911.xml">(1985) 3 Out of 5 Live
      (1985)      What Is This?</t_str>
  </a>
</q>

```

Fig. 3. Excerpt of the English test collection in AVE 2007

Participant systems must return one of the following values for each answer according to the response format (see Figure 4):

- **VALIDATED.** Indicates that the answer is correct and supported by the given text. There is no restriction in the number of **VALIDATED** answers (from zero to all).
- **SELECTED** indicates that the answer is **VALIDATED** and it is the one chosen as the output of a hypothetical QA system. The **SELECTED** answers are evaluated against the QA systems of the Main Track. No more than one answer per question can be marked as **SELECTED**. At least one of the **VALIDATED** answers must be marked as **SELECTED**.
- **REJECTED** indicates that the answer is incorrect or there is not enough evidence of its correctness. There is no restriction in the number of **REJECTED** answers (from zero to all).

This configuration permitted us to compare the AV systems responses with the QA ones, and obtain some evidences about the gain in performance that sophisticated AV modules can give to QA systems (see below).

q_id a_id [SELECTED VALIDATED REJECTED] confidence

Fig. 4. Response format in AVE 2007

3 Collections

Since our objective was to compare AVE results with the QA main track results, we must ensure that we give to AV systems no extra information. The fact of grouping all the answers to the same question could lead to provide extra information based on counting answer redundancies that QA systems might not be considering. For this reason we removed duplicated answers inside the same question group. In fact, if an answer was contained in another answer, the shorter one was removed. Finally, NIL answers, void answers and answers with a supporting snippet larger than 700 characters (maximum permitted in the main track) were discarded for building the collections. This processing lead to a reduction in the number of answers to be validated (see Tables 1 and 2): from 11.2% in the Italian test collection to 88.3% in the Bulgarian development collection.

For the assessments, we reused the QA judgements because they were done considering the supporting snippets in a similar way the AV systems must do. The relation between QA assessments and AVE judgements was the following:

- Answers judged as Correct have a value equal to **VALIDATED**
- Answers judged as Wrong or Unsupported have a value equal to **REJECTED**
- Answers judged as Inexact have a value equal to **UNKNOWN** and are ignored for evaluation purposes.
- Answers not evaluated at the QA main track (if any) are also tagged as **UNKNOWN** and they are also ignored in the evaluation.

3.1 Development Collections

Development collections were obtained from the QA@CLEF 2006 [6] main track questions and answers. Table 1 shows the number of questions and answers for each language together with the percentage that these answers represent over the number of answers initially available, and the number of answers with **VALIDATED** and **REJECTED** values.

These collections were available for participants after their registration at CLEF at <http://nlp.uned.es/QA/ave/>

Table 1. Number of questions and answers in the AVE 2007 development collections

	German	English	Spanish	French	Italian	Dutch	Portuguese	Bulgarian
Questions	187	200	200	200	192	198	200	56
Answers (final)	504	1121	1817	1503	476	528	817	70
% over available answers	31.5	62.28	53.44	50.1	47.6	44	40.85	11.67
VALIDATED	135	130	265	263	86	100	153	49
REJECTED	369	991	1552	1240	390	428	664	21

3.2 Test Collections

Test collections were obtained from the QA@CLEF 2007 main track. In this edition, questions were grouped by topic [4]. The first question of a topic was self contained in the sense that there is no need of information outside the question to answer it. However, the rest of the topic questions can refer to implicit information linked to the previous questions and answers of the topic group (anaphora, co-reference, etc.).

For the AVE 2007 test collections we only made use of the self-contained questions (the first one of each topic group) and their respective answers given by the participant systems in QA.

Table 2. Number of questions and answers in the AVE 2007 test collections¹

	German	English	Spanish	French	Italian	Dutch	Portuguese	Romanian
Questions	113	67	170	122	103	78	149	100
Answers (final)	282	202	564	187	103	202	367	127
% over available answers	48.62	60.3	66.35	75.4	88.79	51.79	30.58	52.05
VALIDATED	67	21	127	85	16	31	148	45
REJECTED	197	174	424	86	84	165	198	58
UNKNOWN	18	7	13	16	3	6	21	24

The change of the task produced a lower participation in the QA main track because systems were not tuned on time and this fact, together with the consideration of less number of questions and the elimination of redundancies led to a reduction of the evaluation corpora in AVE 2007.

Table 2 shows the number of questions and the number of answers to be validated (or rejected) in the test collections together with the percentage that these answers represent over the answers initially available.

4 Evaluation of the Answer Validation Exercise

In [7] was argued why the AVE evaluation is based on the detection of the correct answers. Instead of using an overall accuracy as the evaluation measure, we proposed the use of precision (1), recall (2) and F-measure (3) (harmonic mean) over answers that must be VALIDATED. In other words, we proposed to quantify systems ability to detect whether there is enough evidence to accept an answer.

Results can be compared between systems but always taking as reference the following baselines:

¹ French assessments not available when this report was submitted.

1. A system that accepts all answers (return VALIDATED or SELECTED in 100% of cases)
2. A system that accepts 50% of the answers (random)

$$precision = \frac{|predicted\_correctly\_as\_SELECTED\_or\_VALIDATED|}{|predicted\_as\_SELECTED\_or\_VALIDATED|} \quad (1)$$

$$recall = \frac{|predicted\_correctly\_as\_SELECTED\_or\_VALIDATED|}{|CORRECT\_answers|} \quad (2)$$

$$F = \frac{2 \cdot recall \cdot precision}{recall + precision} \quad (3)$$

However, this is an intrinsic evaluation that is not enough for comparing AVE results with QA results in order to obtain some evidence about the goodness of incorporating more sophisticated validation systems into the QA architecture. Some recent works [5] have shown how the use of textual entailment can improve the accuracy of QA systems. Our aim was to obtain evidences of this improvement in a comparative and shared evaluation.

For this reason, a new measure (4), very easy to understand, was applied in AVE 2007. Since answers were grouped by questions and AV systems were requested to SELECT one or none of them, the resulting behaviour is comparable to a QA system: for each question there is no more than one SELECTED answer. The proportion of correctly selected answers is a measure comparable to the accuracy used in the QA Main Track and, therefore, we can compare AV systems taking as reference the QA systems performance over the questions involved in AVE test collections.

This measure has an upper bound given by the proportion of questions that have at least one correct answer (in its corresponding group). This upper bound corresponds to a perfect selection of the correct answers given by all the QA systems at the main track. The normalization of qa_accuracy with this upper bound is given in (5). We will refer to this measure also as percentage of the perfect selection (normalized_qa_accuracy x 100).

$$qa\_accuracy = \frac{|answers\_SELECTED\_correctly|}{|questions|} \quad (4)$$

$$normalized\_qa\_accuracy = \frac{|answers\_SELECTED\_correctly|}{|questions\_with\_correct\_answers|} \quad (5)$$

$$random\_qa\_accuracy = \frac{1}{|questions|} \sum_{q \in questions} \frac{|correct\_answers\_of(q)|}{|answers\_of(q)|} \quad (6)$$

Besides the upper bound, results of qa_accuracy can be compared with the following baseline system: A system that validates 100% of the answers and selects

randomly one of them. Thus, this baseline can be seen as the average proportion of correct answers per question group (6).

5 Results

Nine groups (2 less than the past edition) have participated in four different languages. Table 3 shows the participant groups and the number of runs they submitted per language. Again, English and Spanish were the most popular with 8 and 5 runs respectively.

Tables 4-7 show the results for all participant systems in each language. Results cannot be compared between languages since the number of answers to be validated and the proportion of the correct ones are different for each language (due to the real submission of the QA systems). Together with the systems precision, recall and F-measure, the two baselines values are shown: the results of a system that always accept all answers (validates 100% of the answers), and the results of a hypothetical system that validates the 50% of answers.

In our opinion, F-measure is an appropriate measure to identify the systems that perform better, measuring their ability to detect the correct answers and only them. However, we wanted to obtain some evidence about the improvement that more sophisticated AV systems could provide to QA systems. Tables 8-11 show the rankings of systems (merging QA and AV systems) according to the QA accuracy calculated only over the subset of questions considered in AVE 2007. With the exception of Portuguese where there is only one participant group, there are AV systems for each language able to achieve more than 70% of the perfect selection. In German and English, the best AV systems obtained better results than the QA systems, achieving a 93% of the perfect selection in the case of German.

Table 3. Participants and runs per language in AVE 2007

	German	English	Spanish	Portuguese	Total
Fernuniversität in Hagen	2				2
U. Évora				1	1
U. Iasi		1			1
DFKI		2			2
INAOE			2		2
U. Alicante		2			2
Text Mess project		2			2
U. Jaén			2		2
UNED		1	1		2
Total	2	8	5	1	16

Table 4. Precision, Recall and F measure over correct answers for Spanish

Group	System	F	Precision	Recall
INAOE	tellez_1	0.53	0.38	0.86
INAOE	tellez_2	0.52	0.41	0.72
UNED	rodrigo	0.47	0.33	0.82
UJA	magc_1	0.37	0.24	0.85
100% VALIDATED		0.37	0.23	1
50% VALIDATED		0.32	0.23	0.5
UJA	magc_2	0.19	0.4	0.13

Table 5. Precision, Recall and F measure over correct answers for German

Group	System	F	Precision	Recall
FUH	iglockner_1	0.72	0.61	0.9
FUH	iglockner_2	0.68	0.54	0.94
100% VALIDATED		0.4	0.25	1
50% VALIDATED		0.34	0.25	0.5

Table 6. Precision, Recall and F measure over correct answers for English

Group	System	F	Precision	Recall
DFKI	ltqa_2	0.55	0.44	0.71
DFKI	ltqa_1	0.46	0.37	0.62
U. Alicante	ofe_1	0.39	0.25	0.81
Text-Mess Project	Text-Mess_1	0.36	0.25	0.62
Iasi	adiftene	0.34	0.21	0.81
UNED	rodrigo	0.34	0.22	0.71
Text-Mess Project	Text-Mess_2	0.34	0.25	0.52
U. Alicante	ofe_2	0.29	0.18	0.81
100% VALIDATED		0.19	0.11	1
50% VALIDATED		0.18	0.11	0.5

Table 7. Precision, Recall and F measure over correct answers for Portuguese

Group	System	F	Precision	Recall
UE	jsaias	0.68	0.91	0.55
100% VALIDATED		0.6	0.43	1
50% VALIDATED		0.46	0.43	0.5

In general, the groups that participated in both QA Main Track and AVE, obtained better results with the AV system than with the QA one. This can be due to two factors: Or they need to extract more and better candidate answers, or they do not use their own AV module to rank them properly in the QA system.

Table 8. Comparing AV systems performance with QA systems in Spanish

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.59	100%
Priberam		QA	0.49	83.17%
INAOE	tellez_1	AV	0.45	75.25%
UNED	rodrigo	AV	0.42	70.3%
UJA	magc_1	AV	0.41	68.32%
INAOE		QA	0.38	63.37%
INAOE	tellez_2	AV	0.36	61.39%
Random		AV	0.25	41.45%
MIRA		QA	0.15	25.74%
UPV		QA	0.13	21.78%
UJA	magc_2	AV	0.08	13.86%
TALP		QA	0.07	11.88%

Table 9. Comparing AV systems performance with QA systems in German

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.54	100%
FUH	iglockner_2	AV	0.50	93.44%
FUH	iglockner_1	AV	0.48	88.52%
DFKI	dfki071dede	QA	0.35	65.57%
FUH	fuha071dede	QA	0.32	59.02%
Random		AV	0.28	51.91%
DFKI	dfki071ende	QA	0.25	45.9%
FUH	fuha072dede	QA	0.21	39.34%
DFKI	dfki071ptde	QA	0.05	9.84%

All the participant groups in AVE 2007 reported the use of an approach based on Textual Entailment. 5 of the 9 groups (FUH, U. Iasi, INAOE, FUH, U. Évora and DFKI) have also participated in the Question Answering Track, showing that techniques developed for Textual Entailment are in the process of being incorporated in the QA systems participating at CLEF.

Table 12 shows the techniques used by AVE participant systems. In general, the groups that performed some kind of syntactic or semantic analysis worked in the Automatic Hypothesis Generation as a combination of the question and the answer. However, in some cases the hypothesis generated was directly in a logic form instead of a textual sentence.

All the participants reported the use of lexical processing. Lemmatization and part of speech tagging were commonly used. In the other side, only few systems used first order logic representations, performed semantic analysis and took the validation decision with a theorem prover.

Table 10. Comparing AV systems performance with QA systems in English

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.3	100%
DFKI	Itqa_2	AV	0.21	70%
Iasi	adiftene	AV	0.21	70%
UA	ofe_2	AV	0.19	65%
U.Indonesia	CSUI_INEN	QA	0.18	60%
UA	ofe_1	AV	0.18	60%
DFKI	Itqa_1	AV	0.16	55%
UNED	rodrigo	AV	0.16	55%
Text-Mess Project	Text-Mess_1	AV	0.15	50%
DFKI	DFKI_DEEN	QA	0.13	45%
Text-Mess Project	Text-Mess_2	AV	0.12	40%
Random		AV	0.1	35%
DFKI	DFKI_ESEN	QA	0.04	15%
Macquarie	MQAF_NLEN_1	QA	0	0%
Macquarie	MQAF_NLEN_2	QA	0	0%

Table 11. Comparing AV systems performance with QA systems in Portuguese

Group	System	System Type	QA accuracy	% of perfect selection
Perfect selection		QA	0.74	100%
Priberam		QA	0.61	82.73%
UE	jsaias	AV	0.44	60%
Random		AV	0.44	60%
U. Evora	diue	QA	0.41	55.45%
LCC	lcc_ENPT	QA	0.3	40%
U. Porto	feup	QA	0.23	30.91%
INESC-ID	CLEF07-2_PT	QA	0.13	17.27%
INESC-ID	CLEF07_PT	QA	0.11	15.45%
SINTEF	esfi_1	QA	0.07	10%
SINTEF	esfi_2	QA	0.04	5.45%

Lexical similarity was the feature most used for taking the validation decision. In general, systems that performed syntactic or semantic processing used this processing as similarity features. None of the systems reported the use of semantic frames.

Table 12. Techniques, resources and methods used by the AVE participants

	adifrene	tellez	rodrigo	iglock-	iglock-	jsaias	ltqa	magc	ofe	text_mess
Generates hypotheses	X	X		X	X				X	X
Wordnet	X			X	X					
Chunking		X				X		X		
n-grams, longest common Subsequences		X					X	X	X	X
Phrase transformations	X	X								
NER	X	X	X					X		X
Num. expressions	X	X	X		X	X				X
Temp. expressions			X		X	X				X
Coreference resolution				X	X					
Dependency analysis	X						X		X	
Syntactic similarity	X	X					X		X	
Functions (sub, obj, etc)	X					X	X			
Syntactic transformations	X									
Word-sense disambiguation				X	X					
Semantic parsing	X			X	X	X				
Semantic role labeling				X	X					
First order logic representation				X	X	X				
Theorem prover				X	X	X				
Semantic similarity	X					X				

6 Conclusions

In this second edition of the Answer Validation Exercise, techniques developed for Recognizing Textual Entailment have been employed widely, although the exercise was defined more closely to the real answer validation application.

We have refined the evaluation methodology in order to consider the QA systems performance as a reference for AV systems evaluation. Thus, new measures have been defined together with their respective baselines: *qa_accuracy* and the percentage of the perfect selection (*normalized_qa_accuracy*).

With respect to the development of test collections, the new evaluation framework led us to reduce redundancies in the sets of answers. This process reduces the size of the testing collections discarding around 50% of candidate answers. The training and testing collections resulting from AVE 2006 and 2007 are available at <http://nlp.uned.es/QA/ave> for researchers registered at CLEF.

Results show that AV systems are able to detect correct answers improving the results of QA systems. In fact, except for Portuguese (where there is only one participant at AVE), all the systems are far from the random behaviour and closer to the perfect selection (from 70% to 93%).

All systems utilize lexical processing, most of them introduce a syntactic level and only few make use of semantics and logic. Groups that participated in both QA and AVE tracks show better performance in the selection of answers than the results obtained by the whole QA system. This fact points to the need of considering the evidences given by the AV modules in order to generate more and better candidate answers. In this way, the approach of looping the AV module with the generation of candidate answers should be considered instead of the solely approach based on the ranking of candidate answers.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology within the Text-Mess-INES project (TIN2006-15265-C06-02), the Education Council of the Regional Government of Madrid and the European Social Fund. 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).

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, Venice, Italy (2006)
2. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) *MLCW 2005. LNCS (LNAI)*, vol. 3944, pp. 177–190. Springer, Heidelberg (2006)
3. Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B.: The Third PASCAL Recognizing Textual Entailment Challenge. In: *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (2007)
4. Giampiccolo, D., et al.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: *Working Notes of CLEF 2007* (2007)
5. Harabagiu, S., Hickl, A.: Methods for Using Textual Entailment in Open-Domain Question Answering. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, pp. 905–912 (2006)
6. 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006. LNCS*, vol. 4730. Springer, Heidelberg (2007)
7. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the Answer Validation Exercise 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006. LNCS*, vol. 4730. Springer, Heidelberg (2007)

Overview of QAST 2007

Jordi Turmo¹, Pere R. Comas¹, Christelle Ayache², Djamel Mostefa²,
Sophie Rosset³, and Lori Lamel³

¹ TALP Research Center, Technical University of Catalonia (UPC)

turmo@lsi.upc.edu, pcomas@lsi.upc.edu

² ELDA/ELRA, Paris, France

ayache@elda.org, mostefa@elda.org

³ LIMS, Paris, France

rosset@limsi.fr, lamel@limsi.fr

Abstract. This paper describes QAST, a pilot track of CLEF 2007 aimed at evaluating the task of Question Answering in Speech Transcripts. The paper summarizes the evaluation framework, the systems that participated and the results achieved. These results have shown that question answering technology can be useful to deal with spontaneous speech transcripts, so for manually transcribed speech as for automatically recognized speech. The loss in accuracy from dealing with manual transcripts to dealing with automatic ones implies that there is room for future research in this area.

Keyword: Question answering, Spontaneous speech transcripts.

1 Introduction

The task of Question Answering (QA) consists of providing short, relevant answers to natural language questions. Most Question Answering research has focused on extracting information from text sources, providing the shortest relevant text in response to a question [4,5]. For example, the correct answer to the question *How many groups participate in the CHIL project?* is 16. Whereas the response to the question of *who are the partners in CHIL?* is a list of the partners. This simple example illustrates the two main advantages of QA has over current search engines: first, the input is a natural language question rather a keyword query, and second, the answer provides the desired information content and not a potentially large set of documents or URLs that the user must plow through.

Most of current QA systems handle independent questions and produce one answer to each question, extracted from textual data, for both open domain and limited domain tasks. However, a large portion of human interactions involve spontaneous speech, e.g. meetings, seminars, lectures, telephone conversations, and are beyond the capacities of current text-based factual QA systems. Most of the recent QA research has been undertaken by natural language groups who have typically applied techniques to written texts, and assume that these texts have a correct syntactic and semantic structure. The grammatical structure of

spoken language is different from that of written language, and some of the anchor points used in text processing such as punctuation must be inferred and are therefore error prone. Other spoken language phenomena include disfluencies, repetitions, restarts and corrections. In the case that automatic processing is used to create the speech transcripts, an additional challenge is dealing with the recognition errors. The lecture and interactive meeting data are particularly difficult due to run-on sentences (where the distance between the first part of an utterance and its end one can be very long) and interruptions. Therefore current techniques for text-based QA need substantial adaptation in order to access the information contained in audio data.

This paper provides an overview of a pilot evaluation track at CLEF 2007 for Question Answering in Speech Transcriptions, named QAST. Section 2 describes the principles of this evaluation track. Sections 3 and 4 present the evaluation framework and the systems that participated, respectively. Section 5 shows the results achieved and the main implications. Finally, Section 6 concludes.

2 The QAST Task

The objective of this pilot track is to provide a framework in which QA systems can be evaluated when the answers have to be found in spontaneous speech transcripts (manual and automatic transcripts). There are three main objectives to this evaluation:

- Comparing the performances of the systems dealing with both types of transcripts.
- Measuring the loss of each system due to the inaccuracies in state of the art ASR technology.
- Motivating and driving the design of novel and robust factual QA architectures for automatic speech transcripts.

In this evaluation, the QA systems have to return answers found in the audio transcripts to questions presented in a written natural language form. The answer is the minimal sequence of words that includes the correct exact answer in the audio stream. For the purposes of this evaluation, instead of pointers in the audio signal, the recognized words covering the location of the exact answer have to be returned. For example, consider the question *which organisation has worked with the University of Karlsruhe on the meeting transcription system?*, and the following extract of an automatically recognized document:

breath fw and this is , joint work between University of Karlsruhe and coming around so fw all sessions , once you find fw like only stringent custom film canals communicates on on fw tongue initials .

corresponding to the following exact manual transcript:

uhm this is joint work between the University of Karlsruhe and Carnegie Mellon, so also here in these files you find uh my colleagues and uh Tanja Schultz.

The answer found in the manual transcript is *Carnegie Mellon* whereas in the automatic transcript it is *coming around*. This example illustrates the two principles that guide this track:

- The questions are generated considering the exact information in the audio stream regardless of how this information is transcribed, because the transcription process is transparent to the user.
- The answer to be extracted is the minimal sequence of words that includes the correct exact answer in the audio stream (i.e., in the manual transcripts). In the above example, the answer to be extracted from the automatic transcript is *coming around*, because this text gives the start/end pointers to the correct answer in the audio stream.

Four tasks have been defined for QAST:

- T1: QA in manual transcriptions of lectures.
- T2: QA in automatic transcriptions of lectures.
- T3: QA in manual transcripts of meetings.
- T4: QA in automatic transcriptions of meetings.

3 Evaluation Protocol

3.1 Data Collections

The data for the QAST pilot track consists of two different resources, one for dealing with the lecture scenario and the other for dealing with the meeting scenario:

- The CHIL corpus¹: it consists of around 25 hours (around 1 hour per lecture) both manually and automatically transcribed (LIMSI produced the ASR transcriptions with around 20% of word error rate -WER- [2], while the manual ones were done by ELDA). In addition, the set of lattices and confidences for each lecture has been provided. The domain of the lectures is *speech and language processing*. The language is European English (mostly spoken by non native speakers). Lectures have been provided with simple tags. Seminars are formatted as plain text files (ISO-8859-1) [3].
- The AMI corpus²: it consists of around 100 hours (168 meetings) both manually and automatically transcribed (the University of Edinburgh produced the ASR transcripts with around 38% of WER [1]). The domain of this meetings is *design of television remote control*. The language is European English. Meetings (as lectures) have been produced with simple tags. Meetings are formatted as plain text files (ISO-8859-1).

¹ <http://chil.server.de>

² <http://www.amiproject.org>

Questions and Answer Types. For each one of the scenarios, two sets of questions will be provided to the participants:

- Development set (1 February 2007) :
 - Lectures: 10 seminars and 50 questions.
 - Meetings: 50 meetings and 50 questions.
- the Evaluation set (18 June 2007):
 - Lectures: 15 seminars and 100 questions.
 - Meetings: 118 meetings and 100 questions.

Question sets have been formatted as plain text files, with one question per line as defined in the Guidelines³. All the questions in the QAST task are factual questions, whose expected answer is a Named Entity (person, location, organization, language, system, method, measure, time, color, shape and material). No definition questions have been proposed. The two data collections (CHIL and AMI corpus) were first tagged with Named Entities. Then, an English native speaker created questions for each NE tagged session. So each answer is a tagged Named Entity.

An answer is basically structured as an [answer-string, document-id] pair, where the answer-string contains nothing more than a complete and exact answer (a Named Entity) and the document-id is the unique identifier of a document that supports the answer. There are no particular restrictions on the length of an answer-string (which is usually very short), but unnecessary pieces of information will be penalised, since the answer will be marked as non-exact. Assessors will focus mainly on the responsiveness and usefulness of the answers.

3.2 Human Judgement

The files submitted by participants have been manually judged by native speaking assessors. Assessors considered correctness and exactness of the returned answers. They have also checked that the document labelled with the returned docid supports the given answer. One assessor evaluated the results. Then, another assessor manually checked each judgement evaluated by the first one. Any doubts about an answer was solved through various discussions.

To evaluate the data, assessors used an evaluation tool developed in Perl (at ELDA) named QASTLE⁴. A simple interface permits easy access of the question, the answer and the document associated with the answer (all in one window only).

For T2 and T4 (QA on automatic transcripts) the manual transcriptions were aligned to the automatic ASR outputs to find the answer in the automatic transcripts. The alignments between the automatic and the manual transcription were done using time information for most of the seminars and meetings. Unfortunately for some AMI meetings time information were not available and only word alignments were used.

³ <http://www.lsi.upc.edu/~qast>

⁴ <http://www.elda.org/qastle/>

After each judgement the submission files have been modified. A new element appears in the first column: the answer's evaluation (or judgement). The four possible judgements (also used at TREC[5]) correspond to a number ranging between 0 and 3:

- 0 correct: the answer-string consists of the relevant information (exact answer), and the answer is supported by the returned document.
- 1 incorrect: the answer-string does not contain a correct answer or the answer is not responsive.
- 2 non-exact: the answer-string contains a correct answer and the docid supports it, but the string has bits of the answer missing or is longer than the required length of the answer.
- 3 unsupported: the answer-string contains a correct answer but the docid does not support it.

3.3 Measures

The two following metrics used in CLEF have been used in the QAST evaluation:

1. Mean Reciprocal Rank (MRR) measures how well ranked is the right answer, as defined in Section 2, in the list of 5 possible answers in average.
2. Accuracy: The fraction of correct answers ranked in the first position in the list of 5 possible answers.

4 Submitted Runs

A total of five groups from five different countries submitted results for one or more of the proposed QAST tasks. Due to various reasons (technical, financial, etc.), three other registered groups were not be able to submit any results.

The five participating groups are the following:

- CLT, Center for Language Technology, Australia;
- DFKI, Germany;
- LIMSI, Laboratoire d'Informatique et de Mécanique des Sciences de l'Ingénieur, France;
- TOKYO, Tokyo Institute of Technology, Japan;
- UPC, Universitat Politècnica de Catalunya, Spain.

Five groups participated in both T1 and T2 tasks (CHIL corpus) and three groups participated in both T3 and T4 tasks (AMI corpus).

The participants could submit up to 2 submissions per task and up to 5 answers per question. The systems used in the submissions are described in Table 1. In total, 28 submissions were evaluated: 8 submissions from 5 participating sites for T1, 9 submission files from 5 different sites for T2, 5 submissions from 3 participants for T3 and 6 submissions from 3 participants for T4. The lattices provided for task T2 were not finally used by any participant.

Table 1. Systems that participated in QAST

System	Enrichment	Question classification	Doc/Pass Retrieval	Answer Extraction	NERC
clt1	words and NEs	hand-crafted patterns	pass. ranking based on word similarities between pass. and query	candidate ranking based on frequency and the NER confidence	hand-crafted patterns, gazeetters and ME models
clt2					No ME models
dfki1	words and NEs	hand-crafted sint.-sem. rules	Lucene	candidate ranking based on frequency	gazeetters and not tuned statistical models
limsi1	words and NEs	hand-crafted patterns	pass. ranking based on hand-crafter back-off queries	candidate ranking based on frequency, keyword distance and retrieval confidence	hand-crafted patterns
limsi2			cascaded doc/pass ranking based on search descriptors		
tokyo1	words	non-linguistic statistical multi-word model	pass. retrieval with interpolated doc/pass statistical models	candidate ranking based on statistical multi-word model	no
tokyo2			addition of word classes to the statistical models		
upc1	words, NEs lemmas and POS	perceptrons	pass. ranking based on iterative query relaxation	candidate ranking based on keyword distance and density	hand-crafted patterns, gazeetters and perceptrons
upc2	also phonetics		addition of approximated phonetic matching		

5 Results

The results for the four QAST tasks are presented in tables 2, 3, 4 and 5. Due to some problems (typo, answer type) some questions have been deleted from the scoring results in tasks T1, T2 and T3. In total, the results have been calculated on the basis of 98 questions for tasks T1 and T2, and 96 for T3. In addition, and due to also missing time information at word level for some AMI meetings, seven questions have been deleted from the scoring results of T4. The results for this task have been calculated on the basis of 93 questions.

The results are very encouraging. First, the best result in accuracy achieved in tasks involving manual transcripts (0.51 for task T1) is closed to the best two results for factual questions in TREC 2006 (0.58 and 0.54), in which monolingual English QA was evaluated. Second, this behaviour is also observed in average: the accuracy in average achieved in tasks T1 and T3 is 0.22, which is comparable with 0.18 achieved in TREC 2006. Although no direct comparisons between QAST and TREC are possible due to the use of different data, questions and answer types, these facts show that QA technology can be useful to deal with spontaneous speech transcripts.

Table 2. Results for T1 (QA on CHIL manual transcriptions)

System	# Questions	#Correct answers	MRR	Accuracy
clt1_t1	98	16	0.09	0.06
clt2_t1	98	16	0.09	0.05
dfki1_t1	98	19	0.17	0.15
limsi1_t1	98	43	0.37	0.32
limsi2_t1	98	56	0.46	0.39
tokyo1_t1	98	32	0.19	0.14
tokyo2_t1	98	34	0.20	0.14
upcl_t1	98	54	0.53	0.51

Table 3. Results for T2 (QA on CHIL automatic transcriptions)

System	#Questions	#Correct answers	MRR	Accuracy
clt1_t2	98	13	0.06	0.03
clt2_t2	98	12	0.05	0.02
dfki1_t2	98	9	0.09	0.09
limsi1_t2	98	28	0.23	0.20
limsi2_t2	98	28	0.24	0.21
tokyo1_t2	98	17	0.12	0.08
tokyo2_t2	98	18	0.12	0.08
upcl_t2	96	37	0.37	0.36
upc2_t2	97	29	0.25	0.24

Table 4. Results for T3 (QA on AMI manual transcriptions). *Due to a bug with the output format script, UPC asked to the assessors to reevaluate their unique run for T3. The results in brackets must be regarded as a non official run.

System	#Questions	#Correct answers	MRR	Accuracy
clt1_t3	96	31	0.23	0.16
clt2_t3	96	29	0.25	0.20
limsi1_t3	96	31	0.28	0.25
limsi2_t3	96	40	0.31	0.25
upcl_t3*	95	23(27)	0.22(0.26)	0.20(0.25)

Table 5. Results for T4 (QA on AMI manual transcriptions)

System	#Questions	#Correct answers	MRR	Accuracy
clt1_t4	93	17	0.10	0.06
clt2_t4	93	19	0.13	0.08
limsi1_t4	93	21	0.19	0.18
limsi2_t4	93	21	0.19	0.17
upcl_t4	91	22	0.22	0.21
upc2_t4	92	17	0.15	0.13

Finally, the accuracy values are 0.22 and 0.15 in average for the tasks involving lectures (T1 and T2, respectively), and 0.21 and 0.14 for those involving meetings (T3 and T4, respectively). These values show that the accuracy decreases in average more than 36% when dealing with automatic transcripts. The reduction of this difference between accuracy values have to be taken as a main goal in the future research.

6 Conclusion

In this paper, we have described the QAST 2007 (Question Answering in Speech Transcripts) task. A set of five groups participated in this track with a total of 28 submitted runs among four specific tasks. In general, the results achieved show that, first, QA technology can be useful to deal with spontaneous speech transcripts, and second, the loss in accuracy when dealing with automatically transcribed speech is high. These results are very encouraging and suggest that there is room for future research in this area.

Future work aims at including in the evaluation framework other languages than English, oral questions, and other question types different than factual ones.

Acknowledgments

We are very grateful to Thomas Hain from the University of Edimburgh, who provide us with the AMI transcripts automatically generated by their ASR. This work has been jointly funded by the European Commission (CHIL project IP-506909), the Spanish Ministry of Science (TEXTMESS project) and the LIMSI AI/ASP RITEL grant.

References

1. Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., Vepa, J., Wan, V.: The ami system for the transcription of meetings. In: *Proceedings of ICASSP 2007* (2007)
2. Lamel, L., Adda, G., Bilinski, E., Gauvain, J.-L.: Transcribing lectures and seminars. In: *Proceedings of Interspeech 2005* (2005)
3. Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S., Tyagi, A., Casas, J., Turmo, J., Cristoforetti, L., Tobia, F., Pnvmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhaven, R., Bernardin, L., Rochet, C.: The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation Journal* (2008)
4. Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.): *CLEF 2006. LNCS, vol. 4730*. Springer, Heidelberg (2007)
5. Voorhees, E.M., Buckland, L.L. (eds.): *The Fifteenth Text Retrieval Conference Proceedings (TREC 2006)* (2006)

Question Answering with Joost at CLEF 2007*

Gosse Bouma, Geert Kloosterman, 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@rug.nl`

Abstract. We describe our system for the monolingual Dutch and multilingual English to Dutch QA tasks. We describe the preprocessing of Wikipedia, inclusion of query expansion in IR, anaphora resolution in follow-up questions, and a question classification module for the multilingual task. Our best runs achieved 25.5% accuracy for the Dutch monolingual task, and 13.5% accuracy for the multilingual task.

1 Introduction

Joost [1] is a question answering system for Dutch. Document collections are parsed by Alpino [2], a wide-coverage dependency parser for Dutch. Answers are extracted by pattern matching over syntactic 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. Joost also contains a component which collects all answers to questions of a specific type (i.e. birthdates) off-line.

Below, we describe our approach to preprocessing of Wikipedia and query expansion for passage retrieval. Next, we present our system for anaphora resolution in follow-up questions and a novel approach to question classification for the multilingual QA. The results of our system and suggestions for further work are discussed in section 4.

2 Preprocessing and Passage Retrieval

Preprocessing Wikipedia. New in this year's CLEF QA tracks was the inclusion of Wikipedia in the corpus. From the XML-version of Dutch Wikipedia provided by the University of Amsterdam¹ we removed material that was irrelevant for our task (i.e. navigation and pictures), and retained only information that is required to segment the text into titles, sections, and lists. The segmentation is used in the IR index. The text in the simplified XML was tokenized,

* 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.

¹ <http://ilps.science.uva.nl/WikiXML/>

split into (4.7M) sentences, and parsed. For off-line answer extraction, we used the patterns as they were developed for the newspaper corpus, with only minor modifications. In particular, we did not try to extract facts from lists, or on the basis of XML-tags.

Query Expansion for Passage Retrieval. We submitted two runs of our system, applying two different settings of the information retrieval (IR) component which is used to retrieve relevant passages for a given question. Common to both settings is the inclusion of linguistic features in the IR index [3]. IR queries are constructed from questions using various features and feature combinations. Keyword weights have been optimized using a genetic algorithm and data from previous CLEF QA tracks.

The second run includes various forms of query expansion. We experimented with two techniques: global methods using fixed lists and local techniques using blind relevance feedback. For the latter we applied an implementation of the Rocchio algorithm for Lucene [4,5]. A maximum of 10 new keywords was suggested, using the top five documents as positive examples. Furthermore, we applied global expansion techniques using several lists of expansion terms: Wikipedia redirects, synonyms automatically extracted from word-aligned parallel corpora (Europarl) using the techniques described in [6] and ISA-relations of named entities extracted from syntactically annotated monolingual corpora.

3 Follow-Up Questions and Multilingual QA

Follow-up Questions. A new feature in the 2007 QA task are follow-up questions. Questions are grouped in topics (identified by topic-ids), consisting of a number of questions. Answering non-initial questions may require information from previous questions or answers to previous questions. The most important aspect of follow-up questions is anaphora resolution. A noun phrase was considered to be anaphoric if it was a personal or impersonal pronoun, a possessive pronoun, a deictic pronoun, an NP introduced by a deictic determiner, or an NP introduced by a definite determiner and not containing any modifiers. Antecedents were restricted to named entities from the first question/answer pair of a topic. The answer was chosen as antecedent if the initial question was one of a limited number of question types which ask for a named entity (i.e. *what is the capital of*, *who wrote/founded/..*, *who is the chair/president/.. of*). In other cases, the first named entity from the question was chosen. We adopted this naive approach mostly because we lacked data to test and evaluate more sophisticated approaches. Note also that quite a few TREC systems limit anaphora resolution to resolving anaphoric expressions to the topic of the question, apparently with reasonable success.

According to our inspection of the best monolingual run, there were 56 questions which required anaphora resolution. For 29 questions (52%), a correct antecedent for an anaphoric expression was found. In 15 cases (27%), a wrong antecedent was given. An important source of errors were cases where the answer to the initial question was correctly chosen as antecedent, but the answer was

wrong. Incorrect antecedents also occurred when the intended antecedent was not (analysed by the parser as) a named entity. 12 cases (21%) were missed altogether by the anaphora module. These are due to the fact that no attempt was made to treat temporal anaphora such as *toen*, *destijds* (*during that moment/period*), and *daarvoor* (*before this date*), to treat locative uses of *er* (*there*).

Question Classification in Multilingual QA. Our system for multilingual QA performs English to Dutch QA. English questions are translated into Dutch using Babelfish/Systran. One problem with this approach is the fact that proper names and concepts are often mistranslated. We tried to reduce the number of errors using Wikipedia. For each name or concept in the English question, we check if it has an English Wikipedia lemma that links to a corresponding Dutch page. If so, the name of the Dutch lemma is used in the translation. Otherwise, the English name is used in the translation. To improve coverage, we also included Wikipedia redirect pages, and the online geographical database *geonames*² for translation of geographical locations.

A second important aspect of QA-systems is question classification. As many automatically generated translations are grammatically poor, parsing may lead to unexpected results, and, as a consequence, question classification is often incorrect or impossible. To remedy this problem, we included a question classifier for English [7], which we ran on the English source questions. We manually constructed a mapping from the question types used for English to the question types used in Joost. Both the (mapped) English question type and the Joost type assigned to the translated question are used to find an answer to the question. Note that source language question classification is used in many multilingual QA systems, but usually the classification is the same as that used by the target language answer extraction components. Experiments on data-sets from previous years showed that inclusion of the (mapped) question class assigned to the English source question leads to a modest improvement.

4 Evaluation and Conclusion

The results from the CLEF evaluation are given in Table 1 (left). Table 1 (right) gives results per question type for the best Dutch monolingual run. For 20 questions no answer was given (i.e. NIL was returned by the system as answer). There are two main reasons for this: mistakes in anaphora resolution, which made it impossible to find documents or answers matching the question and lack of coverage of the question analysis component. Although there were 28 definition questions, only 18 were classified as such by Joost. List questions were an important source of errors.

Definition questions are answered using a relation-table that was created offline. In addition to these, 24 questions were assigned a question type for which a relation-table existed. This number is lower than for previous CLEF tasks.

² www.geonames.org

Table 1. Official CLEF scores for the monolingual Dutch task and the bilingual English to Dutch task (200 questions), with and without Query Expansion (QE) (left) and per question type for the best Dutch monolingual run (right). A= accuracy (%), R (X,U,W) = number of right (inexact, unsupported, wrong) answers.

Run	A	R	X	U	W	Q type	# q's	A	R	X	U	W
Dutch-mono	24.5	49	11	4	136	Factoids	156	25.6	40	5	4	107
Dutch-mono + QE	25.5	51	10	4	135	List	16	6.3	1	0	5	10
En-Du	13.0	26	8	7	159	Definition	28	35.7	10	0	0	18
En-Du + QE	13.5	27	7	5	161	Temporally	41	19.5	8	3	3	27
						Restricted						
						NIL	20	0.0	0	0	0	20

The impact of adding Wikipedia to the document collection was significant. Although the text version of the Dutch Wikipedia is smaller than the newspaper text collection (approximately 50M and 80M words respectively), 150 of the 180 questions that received an answer were answered using Wikipedia. The inclusion of Wikipedia in the CLEF QA-task has made the task more realistic and attractive. We believe that performance on this task can be improved by taking the structure of Wikipedia more seriously, and by developing methods for relation and answer extraction that combine NLP with XML-based extraction.

Follow-up questions required the incorporation of an anaphora resolution component for questions. The current version of this module performs reasonably well, but its coverage should be extended (to cover locative anaphors and multiple anaphors).

References

1. Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedeman, J.: Linguistic knowledge and question answering. *Traitement Automatique des Langues* 2(46), 15–39 (2005)
2. Bouma, G., van Noord, G., Malouf, R.: Alpino: Wide-coverage computational analysis of Dutch. In: *Computational Linguistics in The Netherlands 2000*, Rodopi, Amsterdam (2001)
3. Tiedemann, J.: Improving passage retrieval in question answering using NLP. In: Bento, C., Cardoso, A., Dias, G. (eds.) *EPIA 2005. LNCS (LNAI)*, vol. 3808, pp. 634–646. Springer, Heidelberg (2005)
4. Rubens, N.: Lucqe - lucene query expansion (2007), <http://lucene-qe.sourceforge.net/>
5. Rubens, N.: The application of fuzzy logic to the construction of the ranking function of information retrieval systems. *Computer Modelling and New Technologies* 10(1), 20–27 (2006)
6. van der Plas, L., Tiedemann, J.: Finding synonyms using automatic word alignment and measures of distributional similarity. In: *Proceedings of ACL/Coling (2006)*
7. Hacıoglu, K., Ward, W.: Question classification with support vector machines and error correcting codes. In: *Proceedings of HLT-NACCL 2003*, Edmonton, Alberta, Canada, pp. 28–30 (2003)

What Happened to Esfinge in 2007?

Luís Miguel Cabral, Luís Fernando Costa, and Diana Santos

Linguatca, Oslo node, SINTEF ICT, Norway
{luis.m.cabral,luis.costa,Diana.Santos}@sintef.no

Abstract. Esfinge is a general domain Portuguese question answering system which uses the information available on the Web as an additional resource when searching for answers. Other external resources and tools used are a broad coverage parser, a morphological analyser, a named entity recognizer and a Web-based database of word co-occurrences.

In this fourth participation in CLEF, in addition to the new challenges posed by the organization (topics and anaphors in questions and the use of Wikipedia to search and support answers), we experimented with a multiple question and multiple answer approach in QA.

Keywords: Question answering, Portuguese, anaphor resolution, question reformulation, answer choice, Wikipedia processing.

1 Introduction

This year's evaluation contest required the systems to adapt to two brand-new conditions: The difficulty of questions was raised by the introduction of topics and anaphoric reference between questions on the same topic; and the difficulty of answers was raised because collections included Wikipedia, in addition to the old newspaper collections. Our main goal this year was therefore to adapt Esfinge to work in these new conditions, which basically consisted in creating an initial module for creating non-anaphoric questions (resolving co-reference) to be input to (the previous year's) Esfinge, and a final module that dealt with the choice of multiple answers from several different collections and/or Esfinge invocations (multi-stream QA). As will be explained below, unexpected problems led us to also try a radically different approach based on a set of patterns obtained from the initial module.

2 Esfinge in 2007

Esfinge participated at CLEF in 2004, 2005 and 2006, as described in detail in the corresponding proceedings. Most work in Esfinge this year was related to address the new challenges introduces in QA@CLEF. Figure 1 gives a general overview of the system used this year.

There is a new **Anaphor Resolution** module to resolve anaphors, which adds, to the original question, a list of alternative questions where the anaphors are (hopefully) resolved. In addition, it may also propose relatively trivial reformulations. Then, for each of the alternative questions:

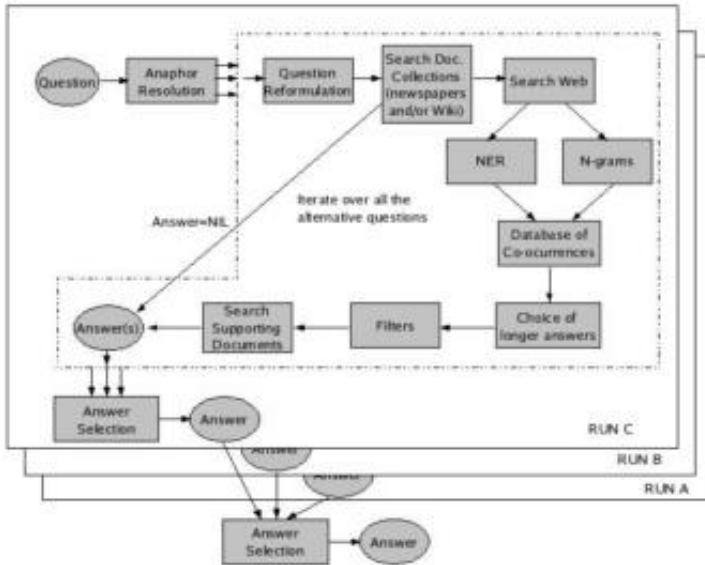


Fig. 1. Architecture of Esfinge 2007

1. The **Question Reformulation** module transforms the question into patterns of plausible answers. These patterns are then searched in the document collection using the **Search Document Collections** module. This module was adapted to allow search also in Wikipedia.
2. If the patterns are not found in the document collections, the system returns NIL and stops. Optionally, it can proceed by searching the same patterns in the Web. Then, all texts retrieved are analysed using a named entity recognizer (NER) system 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 BACO database of co-occurrences [1] and the candidate answers (by ranking order) are analysed in order to check whether they pass a set of filters and to find a document in the collections which supports them.
3. From the moment Esfinge finds a possible answer for the question, it checks only candidate answers that include one of the previously found answers. It will replace the original answer if the new one includes the original answer, passes the filters and has documents in the collection that support it.

After iterating over all alternative questions, Esfinge has a set of possible answers. That is when the new module **Answer Selection** comes to play. This module attempts to select the best answer to the given question, which will be the final answer returned.

3 Anaphor Resolution

We developed a module relying crucially on the PALAVRAS parser [2] to replace anaphoric expressions into fully descriptive expressions (i.e., independently understandable questions).

This question reformulation is our first instantiation of the M,N-O,P model introduced in [3]. Basically, from the input question, we produce an (ordered) set of questions to be input to the original (one question, one answer) Esfinge system. Although this model does not cover everything required by interactive question answering, especially when user follow-up questions relate to previous answers and not to previous questions [4], question reformulation and choice among many answers was high on our research agenda.

The linguistic description of the kinds of anaphors catered for by our system can be found in [5], where we analyse the four sets of 200 questions which had Portuguese as source language. Incidentally, there were quite different kinds of questions depending on the (original source) language, suggesting that more attention should be paid to language differences [6].

In short, we deal rather successfully with (i) pronominal anaphor (subject, direct object, indirect object and oblique); (ii) possessive anaphor; (iii) demonstrative anaphor; and (iv) null subject anaphor; but completely missed, or had bad results, for (v) definite description anaphor; (vi) implicit anaphor; (vii) short questions (incidentally, only in the Portuguese monolingual set); and (viii) anaphoric reference to previous answers. Numbers and examples are in table 1.

PALAVRAS is a broad-coverage dependency parser for Portuguese which is used extensively by Linguateca projects since 1999, resulting in a set of programs to deal with its output described at the AC/DC project website.¹

For anaphor resolution, our hypothesis was that most anaphoric related antecedents would be major constituents. (In fact, this was not confirmed by the data.) So, we set out using PALAVRAS for obtaining argument phrases.

By considering the particular question set, however, it soon became apparent that syntax alone was often not enough to assign the right argument structure. (See again [5] for details.) The simpler the questions, the less syntax is going to help. We have therefore used a set of heuristics – both prior to and after invoking PALAVRAS – to provide for more than one question formulation, to cope with these possible shortcomings.

For each question submitted to PALAVRAS, we get: (i) the anaphoric element and the phrase it is included in, and (ii) a list of possible candidates: all arguments mentioned within the same topic that include a proper name, all adjuncts with the same property, and all proper names and dates as well.

Anaphor resolution proper then proceeds by creating a set of new questions replacing the anaphor with all possible referent candidates. Often, no syntactic clue can help choose which candidate is most appropriate, as in *Quais eram os primeiros nomes dos dois irmãos Piccard? Qual deles ...* or *Qual o período*

¹ <http://www.linguateca.pt/ACDC/>

Table 1. Distribution of the several kinds of anaphors in the material: in parentheses is the subset which depends on the previous answer(s)

Kind	Example question	PT-PT	PT-DE	PT-ES	PT-FR	Total
subject pronoun	Quem é o dono <u>delas</u> ? Quem era <u>ele</u> ?	14	19 (1)	6 (1)	19 (1)	58
personal pronoun	Quem é que <u>o</u> afundou em 1985?	1	1	1	0	3
demonstrative pronoun	Que (...) ao EEE quando <u>este</u> entrou em vigor? Quem é que dirige <u>essa</u> agência?	2 (1)	0	8 (1)	13 (1)	23
possessive pronoun	Qual era o <u>seu</u> verdadeiro nome?	7 (1)	3	4	6	20
null subject	Quantos habitantes tinha?	11 (1)	1	6	3	21
definite desc.	Quantos lugares tem <u>o</u> estádio?	4	7	3	5 (2)	19
implicit	Quem é o actor principal?	6	0	1	2 (1)	9
short questions	Onde?	6	0	0	0	6
other	cada, null object	1	0	1	0	2
Total		52	31	30	48	161

de gestação do ocapí? Qual o seu peso? where one would have to list all three possible noun phrases to get one reformulation right.

We assessed the performance of the anaphoric resolution module in detail, this time, differently from [5], considering also the cases which we had not considered during development. Table 2² provides the system evaluation, as opposed to algorithm evaluation, see [7] for the distinction. It is interesting to note that the Portuguese-only material was the hardest by far.

Table 2. Anaphor resolution performance for the 161 cases (158 questions)

	Number of questions	Correctly detected	Spurious	Undetected	Correctly resolved	Accuracy (resolved/all)
PT-PT	52 (51)	34 (33)	1	18	26 (25)	26/52 (50%)
PT-ES	30	24	2	7	17	17/30 (57%)
PT-DE	31 (29)	22 (21)	5	9	21	21/31 (68%)
PT-FR	48	38	2	10	31	31/48 (64%)
Total	161 (158)	118 (116)	10	44	95	95/161 (59%)

A by-product of the **AnaphorResolution** module was the identification, for each question, of the main verb, its arguments and its adjuncts, together with the possible entities for cross-reference coming from previous analyses inside the same topic. During the submission process, we decided to experiment also with this set of patterns (obtained from syntactic analysis) as an alternative to the original Esfinge patterns. These are called “PALAVRAS patterns” in the present paper. However, since no ranking algorithm was associated to them, their use has to be investigated further to discover how to employ them more judiciously.

² Three questions in the material had two different anaphors.

4 Searching Wikipedia

The use of Wikipedia presented a new challenge for Esfinge. Fearing that the size of the text would make the current methods prohibitively slow (the initial downloaded size amounted to about 5.4G), we chose to store the text in a MySQL database, instead of compiling the text in the IMS-CWB. The process was similar to the one used in BACO, using indexing capabilities to allow faster queries on the collection, indexing words up to a minimum length of 3 characters, and storing the text in sets of several sentences instead of storing the entire article together. In order to keep the sentences' context, information was repeated, intercalating the sentences, instead of simply grouping consecutive sentences, as shown in table 4 of [5].

Having completed the preparation of the data for analysis, the next step consisted in making this data accessible to Esfinge, which was easy, given that Esfinge already used BACO's interface to MySQL that assessed rarity of words, as detailed in [8].

The main task was to make the Wikipedia collection work as just one more resource from which answers could be retrieved, independently of the implementation.

Esfinge generates several text patterns from the given question. Each one is then used to search within the collections. While Esfinge, previously, only catered for CQP patterns to be directly applied to the newspaper collections, corresponding patterns for the MySQL function `Match Against` had to be created to access the indexed text of Wikipedia.

While in CQP Esfinge produces several queries from one expression and later joins the results, in MySQL this was transformed into one single query, independent of word order. For example, the expression *+navegação +cabotagem* matches against the following sentence: *A **cabotagem** se contrapõe à **navegação** de longo curso....*

5 Choosing among Several Answers

For each question reformulation we had one answer, therefore the **Answer selection** module had to choose the final one. Also, we created a large number of runs with different options, employing different search patterns and using different textual resources. Initially we had run the following runs:

- One run with all collections (Web+News+Wiki),
- One run without consulting the Web (News+Wiki),
- One run without the Wikipedia collection (Web+News).

Later we ran the same options but used instead the patterns generated using PALAVRAS.

As we had only two possible runs to send, we used this module also to merge the results of the individual runs. We merged all runs that used the same kind of search patterns (Esfinge or PALAVRAS).

Merging took into consideration the sum of the following aspects: (i) the number of times a certain answer was found in all runs; and (ii) the relevance of the support text to the question asked, computed as the number of times that words (with 3 or more characters) in the question occurred in the support text.

To evaluate this module, we looked into the 378 cases (distributed over the 3 automatic selection runs, presented in Figure 2, as no. 6, 11 and 13) where the choice module had more than one non-NIL answer to choose from, and counted the cases where the right answer was among the candidates (80). For this number, the choice was right in 68.75% of the cases.

6 Our Participation and Additional Experiments

The official results can be seen in the first two lines of Figure 2, together with their subsequent repetition, after several severe bugs were discovered – unfortunately too late to resubmit to CLEF. Figure 2 displays the results of the individual runs and of their combination.

In order to assess the import of the **Answer Selection** module, we did a manual choice run as well (choosing manually among the different answers). This is indicated as **best** selection vs. **automatic** selection.

In order to evaluate the impact of adding Wikipedia as an additional source of knowledge, we also ran last year’s questions with the new architecture (2006A and 2006B, respectively with Esfinge or PALAVRAS patterns), which resulted in a 3-4% improvement only.

Table 3 summarizes the main causes for errors in the best individual run (no. 8). The two main causes for wrong answers are both related to the retrieval of relevant documents. The category “Wrong or incomplete search patterns” refers to questions where the search patterns did not include the necessary information to answer the questions, while “Document retrieval failure” counts the cases

N	Description	Right Answers (Questions)			Partially Right Answers	Invalid Answers	Invalid Answers	Right Answers including NIL			Total NIL
		C	B	A				C	B	A	
		1	2	3				1	2	3	
1	PROBETESTER (Official)	12	0	8	5	2	0	12	0	2	14
2	PROBETESTER (Official)	12	0	10	2	0	0	2	0	5	10
2006A	2006A - Term 1 (W1)	43	2	6	2	0	0	22	2	3	24
	1000 - W1	25	10	3	1	0	0	21	0	5	24
	2006A - Term 2	24	1	1	1	0	0	19	1	6	10
	2006A - Selection 1 (S)	21	1	0	0	0	0	20	1	0	10
	2006A - Selection 2 (S)	40	2	0	0	0	0	25	2	0	10
	2006A - Term 1 (W1)	16	1	0	1	0	0	14	1	0	10
	2006A - W1	25	2	2	0	0	0	19	2	4	10
	1000 - Term 2	22	0	2	1	0	0	21	0	3	10
	2006A - Selection 1 (S)	34	2	0	0	0	0	29	2	0	10
	2006A - Selection 2 (S)	10	0	0	0	0	0	10	0	0	10
2006B	2006B - Selection 1 (S)	34	1	4	2	0	1	24	1	4	10
	2006B - Selection 2 (S)	41	2	0	0	0	0	18	2	0	10
	2006B - Term 1 (W1)	22	0	0	0	0	0	22	0	0	10
	2006B - Term 2	22	0	0	0	0	0	22	0	0	10
	2006B - Term 3	22	0	0	0	0	0	22	0	0	10

Fig. 2. Results of the additional experiments (A: Right answers including NIL; B: Partially right answers on lists; C: Right NIL answers)

where no relevant documents were retrieved in the collections, even though the search patterns included the necessary information. “Other” covers all causes that occurred less than five times.

In this table we counted the first module to fail. This explains why the initial modules are the ones with more errors: the modules which appear later are not even invoked for a significant part of the questions. Even though incompleteness of the search patterns was the main single cause for failure, this was to some extent due to poor communication among some modules (that was discovered only afterwards). It is important to point out that, still, the best run obtained by Esfinge used the PALAVRAS patterns.

7 Discussion and Further Work

We believe that the comparison of Esfinge results in 2006 and 2007 lends support to the claim that this year the difficulty of questions was raised, and we welcome this. Having the questions grouped in topics and including several types of anaphors brings us a step closer to the way humans ask questions and allowed us to develop Esfinge towards higher usefulness.

However, we think that the question set had too many errors to be used as a fair evaluation resource, and we hope that this won’t be repeated in future editions of QA@CLEF.

This year, we concentrated mainly on developing the anaphor resolution module and the module responsible for merging and/or choosing from several alternative answers.

There is a lot of improvement that we can foresee for the first module, although a specific analysis of what errors are due to PALAVRAS performance as opposed to anaphor resolution proper is still due.

The choice algorithm also deserves closer attention, since it attained only 67% or 69% of the best combination when merging 3 runs, and 55% when it tried to merge all runs, producing in fact worse results than some of the individual runs it combined.

To deal with this, we are currently investigating several strategies: (i) to give different weights to different sources, (ii) combine the individual weights that

Table 3. Causes for wrong answers in the best individual run

Wrong	Answers
Co-reference resolution	25
Wrong or incomplete search patterns	63
Document retrieval failure	33
Mistake of the answer scoring algorithm	24
Mistake in the supported answer filter	7
Other	13
Total	165

had been assigned in each individual run, and/or (iii) saving more information, such as the patterns used to find each answer, for aiding the decision.

Acknowledgements. This work was done in the scope of the Linguatca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

References

1. Sarmiento, L.: BACO - A large database of text and co-occurrences. In: Calzolari, N., et al. (eds.) *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 22-28 May 2006, pp. 1787–1790 (2006)
2. Bick, E.: *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press (2000)
3. Santos, D., Costa, L.: QoLA: fostering collaboration within QA. In: Peters, C., et al. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 569–578. Springer, Heidelberg (2007)
4. Bertomeu, N., Uszkoreit, H., Frank, A., Krieger, H.U., Jörg, B.: Contextual phenomena and thematic relations in database QA dialogues: results from a Wizard-of-Oz Experiment. In: *Proceedings of the HLT-NAACL 2006 Workshop on Interactive Question Answering* (2006)
5. Cabral, L.M., Costa, L.F., Santos, D.: Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach. In: Nardi, A., Peters, C. (eds.) *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September (2007)
6. Santos, D., Cardoso, N.: Portuguese at CLEF 2005: Reflections and Challenges. In: Peters, C., ed.: *Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop*, Vienna, Austria, 21-23 September (2005)
7. Mitkov, R.: Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In: *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000)*, Lancaster, UK, pp. 96–107 (2000)
8. Costa, L.: Question answering beyond CLEF document collections. In: Peters, C., et al. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 405–414. Springer, Heidelberg (2007)

Coreference Resolution for Questions and Answer Merging by Validation

Sven Hartrumpf, Ingo Glöckner, and Johannes Leveling

Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen), 58084 Hagen, Germany

Abstract. For its fourth participation at QA@CLEF, the German question answering (QA) system InSicht was improved for CLEF 2007 in the following main areas: questions containing pronominal or nominal anaphors are treated by a coreference resolver; the shallow QA methods are improved; and a specialized module is added for answer merging. Results showed a performance drop compared to last year mainly due to problems in handling the newly added Wikipedia corpus. However, dialog treatment by coreference resolution delivered very accurate results so that follow-up questions can be handled similarly to isolated questions.

1 Overview

Research described in this paper is part of IRSAW¹, a question answering (QA) framework integrating modules for natural language analysis, for combining answer streams, and for logical answer validation (MAVE). Three approaches produce answer candidates from the two corpora for QA@CLEF 2007 (CLEF-News and Wikipedia), resulting in six answer streams, which are merged by MAVE.

The first answer producer is InSicht, a precision-oriented QA system using a semantic network representation of questions and documents (see Sect. 2). The other two answer producers are QAP (Question Answering by Pattern matching) and MIRA (Modified Information Retrieval Approach for QA), see Sect. 3. They employ shallow NLP methods and aim at a high recall to provide a fallback strategy for InSicht. An earlier, but more detailed version of this paper is [1].

2 Changes of InSicht for QA@CLEF 2007

In 2007, the QA@CLEF task was considerably changed both in size and diversity. Adding Wikipedia increased the German corpus from 5 million to 17 million sentences. Moreover, follow-up questions were added. For document processing in InSicht, all documents are parsed by WOCADI (WORD CLASS based DISambiguating parser, [2]) and intratextual coreferences are resolved by CORUDIS [2]. Since the time between guideline and test set release was too short to process the

¹ IRSAW (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web; LIS 4 – 554975(2) Hagen, BIB 48 HGfu 02-01) is funded by the DFG.

Wikipedia, we worked with an older Wikipedia parse. Though only two months older than the official snapshot, it turned out to be considerably different.

In the years before 2007, all questions could be answered in isolation without any reference to the context, like previous questions or answers. The guidelines for QA@CLEF 2007 allowed coreferences *to the topic expressed in the first question/answer pair*. To treat such context-dependent questions, the basic idea was to keep a dialog history containing semantic representations of questions and answers. The dialog history is initialized (i.e. deleted), if the start of a new topic is encountered in the test set. On these semantic representations in the dialog history, coreferences are resolved by the general coreference resolver CORUDIS. This module has already been used successfully on the QA documents.

CORUDIS is a hybrid coreference resolver: it contains symbolic, linguistically motivated *coreference rules* that license possible coreferences and a statistical *multi-dimensional back-off model* derived from a manually annotated corpus for selecting among licensed alternatives. CORUDIS further employs bonus factors for syntactic parallelism, semantic parallelism, and maximality of noun phrases.

29 questions of the 200 German questions of QA@CLEF 2007 require coreference resolution to find an answer (this is only 34.5% of all 84 follow-up questions for the 47 topics with more than one question). Two questions (046, 107) contain an anaphor that corefers with an answer; therefore, answers should also be antecedent candidates. To handle references to non-first questions, we adapted the dialog processing as follows: a subsequent question is deleted from the dialog history only if it contains an anaphor which was successfully resolved. The answer producers used only representations where coreferences had been resolved; questions are rewritten in a form that incorporates all necessary context.

Related work for coreference resolution on the question or document side is surveyed in [3]. Some systems that employ coreference resolution on questions in the Context Task of TREC-10 are described by [4,5]. These approaches utilize coreferences by copying keywords from the question (or answer) containing the antecedent to the question containing the anaphor.

3 Shallow QA Subsystems

QAP [6] employs pattern matching on a per-sentence basis. QAP was improved by adding different classes of questions and more training material for pattern extraction. Several large resources were utilized to create question-answer pairs for training, including data from the authority-controlled PND (*Personennamendatei*) as used in the German Wikipedia. A PND entry contains biographical data about a person such as place and date of birth, place and date of death, aliases, and profession. This data can be transformed to represent question-answer pairs for training pattern extraction. In addition to explicit information, further question-answer pairs are derived, e.g. the age at death.

MIRA [7] is a recall-oriented approach based on IR combined with the selection of the most frequent word sequence tagged with the expected answer type (EAT). The basic method in MIRA is to assign an EAT to the question and process all documents (sentences) presumed to be relevant. The tokens in sentences

are categorized according to the EATs. Answer candidates are then selected by choosing the most frequent word sequences tagged with the EAT.

Due to time constraints, patterns for the Wikipedia data were not produced in time by the shallow QA methods. Instead, the patterns created from CLEF-News were utilized for Wikipedia documents as well.

4 Answer Selection by Logical Validation

The three answer producers delivered one answer stream per corpus to be merged by a new component, the answer validator MAVE [8,9]. The system accepts streams of validation items composed of the question string, the answer string, and a supporting witness text extracted from the document collection. It uses deep linguistic processing and logical reasoning for validating the correctness of answers, i.e. by checking if they are verified by the witness texts.

In order to gain more robustness, the theorem prover of MAVE is embedded in a feedback loop which skips literals until a proof of the reduced set of query literals succeeds. The number of skipped literals serves as a robust indicator for (non)entailment. The system is backed up with tests for false positives which reject trivial or circular answers.

The version of MAVE used for filtering the QA@CLEF 2007 results was mostly identical to the system described by [8], with two main changes: First, extraction of a threshold which makes it possible to reject rather than select the best answer candidate if the evidence is still too weak. Second, integration of large lexical-semantic resources (like GermaNet and OpenThesaurus) which allow more flexible inferences. The current state of the system is detailed by [9].

5 Evaluation and Discussion

We submitted two runs for the German monolingual task. The first run was generated from all six answer streams by applying MAVE for answer selection, while the second run was compiled from QAP and MIRA only in order to obtain a baseline from the shallow QA subsystems. The results (48 right answers (30 in the second run), 2 unsupported ones, 4 inexact ones) dropped in comparison to previous years, mainly because of the addition of Wikipedia and several problems in adapting system components. However, the shallow QA subsystems managed to back up the performance of the deep QA system (18 additional correct answers; the same number of answers were found by InSicht only).

InSicht was able to deal with the extended document collection, but unfortunately the quite unrestricted form of article names led to an inconsistent concept index that rendered many Wikipedia articles inaccessible to InSicht. So, InSicht's answers came too rarely from Wikipedia, which was the main reason for the performance drop. Aggravating this situation, around 50% of the test set questions target Wikipedia documents only. Fortunately, the performance drop in InSicht was in part compensated by the improved shallow QA subsystems and the newly integrated answer validator MAVE.

Compared to positive K1 values in previous years, our system somewhat lost the ability to judge its own answers by assigning accurate scores. This effect was partly due to bugs in the answer validator which blocked the application of important axioms and spoiled results for COUNT and MEASURE questions. Moreover, MAVE was not yet fitted to the modified document collection of QA@CLEF 2007. The dialog treatment was very successful: 89.6% of the questions with anaphors were correctly treated by the coreference resolver.

6 Conclusion

Our system showed a performance drop compared to 2004, 2005, and 2006. Error analysis hinted at the massive change in the size and type of the document collection caused by the addition of Wikipedia. On the positive side, the system architecture matured by the integration of two shallow QA subsystems beside the main, deep QA system, InSicht, and a dedicated answer validator was added, which simplifies the construction of individual streams. In the future, the document processing and the answer producers should be better adjusted to Wikipedia. The successful dialog handling should be tested on more diverse discourse dependency types and structures linking questions and answers.

References

1. Hartrumpf, S., Glöckner, I., Leveling, J.: University of Hagen at QA@CLEF 2007: Coreference resolution for questions and answer merging. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
2. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück (2003)
3. Vicedo, J.L., Ferrández, A.: Coreference in Q & A. In: Strzalkowski, T., Harabagiu, S. (eds.) *Advances in Open Domain Question Answering. Text, Speech and Language Technology*, vol. 32, pp. 71–96. Springer, Dordrecht (2006)
4. Harabagiu, S., Moldovan, D., Paşca, M., Surdeanu, M., Mihalcea, R., Gîrju, R., Rus, V., Lăcătuşu, F., Morărescu, P., Bunesu, R.: Answering complex, list and context questions with LCC's question-answering server. In: Voorhees, E.M., Harman, D. (eds.) *Proceedings of TREC-10*, pp. 355–361 (2001)
5. Oh, J.H., Lee, K.S., Chang, D.S., Seo, C.W., Choi, K.S.: TREC-10 experiments at KAIST: Batch filtering and question answering. In: Voorhees, E.M., Harman, D. (eds.) *Proceedings of TREC-10*, pp. 347–354 (2001)
6. Leveling, J.: On the role of information retrieval in the question answering system IRSAW. In: *Proceedings of LWA 2006, Workshop FGIR*, Hildesheim, Germany, pp. 119–125 (2006)
7. Leveling, J.: A modified information retrieval approach to produce answer candidates for question answering. In: Hinneburg, A. (ed.) *Proceedings of LWA 2007, Workshop FGIR. Gesellschaft für Informatik, Halle/Saale, Germany* (2007)
8. Glöckner, I., Hartrumpf, S., Leveling, J.: Logical validation, answer merging and witness selection – a case study in multi-stream question answering. In: *Proceedings of RIAO 2007, Pittsburgh, USA* (2007)
9. Glöckner, I.: University of Hagen at QA@CLEF 2007: Answer validation exercise. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)

Multilingual Question Answering through Intermediate Translation: LCC's PowerAnswer at QA@CLEF 2007

Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Thomas d'Silva,
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 QA@CLEF 2007 preparation, participation and results. For this exercise, LCC integrated its open-domain PowerAnswer Question Answering system with its statistical Machine Translation engine. For 2007, LCC participated in the English-to-French and English-to-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 improved system and methodology and updates from QA@CLEF 2006.

1 Introduction

In 2006, Language Computer Corporation's open-domain question answering system PowerAnswer [6] participated in QA@CLEF for the first time [1], 2007 is a continuation of this exercise. PowerAnswer has previously participated in many other evaluations, notably NIST's TREC [7] workshop series, however, QA@CLEF is the first Multilingual QA evaluation the system has entered. Additionally, LCC has developed its own statistical machine translation system, which is 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 approach chosen.

The goals for participating in QA@CLEF are (1) to examine how well the QA system performs when given noisy data, such as that from automatic translation and (2) to examine and evaluate the performance and utility of the machine translation system in a question answering environment. To that end, LCC has 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 2007, and then

follows with a discussion of results and challenges in the CLEF question topics. For 2007, LCC participated in the following bilingual tasks: English $\rightarrow$ French, and 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 2007.

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.

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 [13].

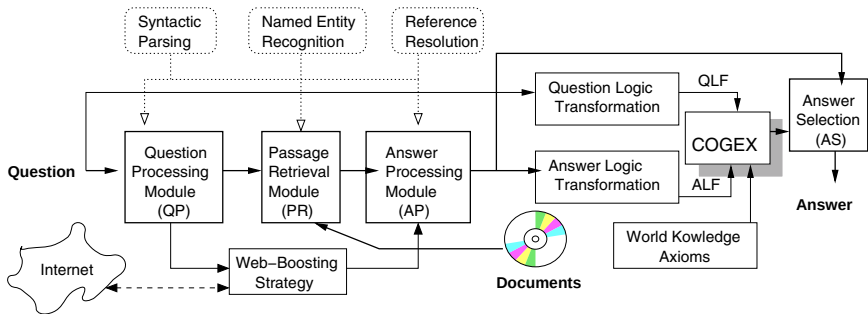


Fig. 1. PowerAnswer 2 Architecture

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 [10]. For QA@CLEF, however, the logic prover is disabled in order to better evaluate the individual components of the QA architecture. COGEX's evaluation on multilingual data was performed in the 2006 CLEF Answer Validation Exercise [16], where the system was the top performer in both Spanish and English.

3 Overview of Translation Engine

The translation system used at LCC – MeTRe – implements phrase-based statistical machine translation [3]; the core translation engine is the open-source Phramer [15] system, developed by one of LCC's engineers. Phramer in turn implements and extends the phrase-based machine translation algorithms described by Koehn [3]. A more detailed description of the MT solution adopted for Multilingual QA@CLEF can be found in [14]. The translation system is trained using the European Parliament Proceedings Parallel Corpus 1996–2003 (EUROPARL) [4], 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 [5] 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 or Portuguese), 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.

¹ <http://www.iccs.inf.ed.ac.uk/~pkoehn/training.tgz>

The capitalization process follows the translation process and it restores the casing of the words, due to using models trained on lowercase text. The capitalization tool uses three-gram statistics extracted from 150 million words from the English GigaWord Second Edition² corpus, augmented with two heuristics:

1. First word will always be uppercased;
2. If the words appear also in the foreign documents, the casing is preserved (this rule is very effective for proper nouns and named entities)

4 PowerAnswer-MeTRe 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). While this fallback method provides excellent usability in a real-world situation, as discussed in the Errors discussion, the method produces answers judged *inexact* in an evaluation framework.

4.1 Passage Retrieval

Making use of PowerAnswer's modular design, for last year's QA@CLEF, 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, and for 2007, the only method used. LCC selected this as the sole method this year because it gave the best performance in terms of quality versus runtime effort. Moreover, LCC has improved the speed of the automatic translator since the 2006 QA@CLEF. In addition to an algorithmic speed improvement of over 100% per execution core, and a decrease in the impact of network latency, the translator also now takes advantage of multiple processors, greatly increasing the time performance of the system. On dual-core machines, the translation speedup is more than 300%.

The entire target language document collection is translated into English, processed through the set of LCC's 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

² <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T12>

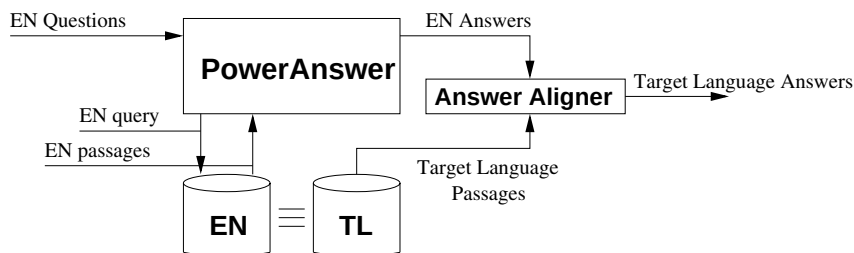


Fig. 2. Passage Retrieval on English documents (default)

one improves the quality of the translation. For 2007, we created all new indexes of the collection. 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. As a main source of errors last year, for 2007 LCC made improvements to the Answer Aligner as described in Section 5. The second solution, as seen in Figure 3, requires minimum effort during 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-based³ IR system) with alternations as the new IR query (step 1). The translation of keywords is performed using MeTRe, 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 MeTRe (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”*). Thus, by re-queried the translated sub-collection (that contains only the translation of those 85 documents) the system retrieves only 72 English documents that will 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.

³ <http://lucene.apache.org/>

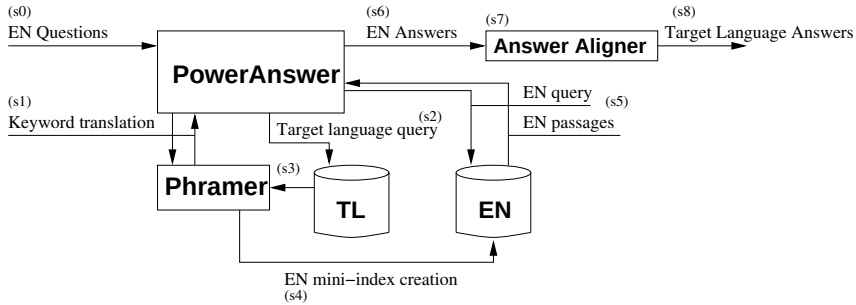


Fig. 3. Passage Retrieval on Target Language documents

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 ([5]). 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.

Table 1. LCC's QA@CLEF 2007 Overall Results

Source	Accuracy	CWS	Improv. from 2006
French	40.50%	0.22234	92.40%
Portuguese	28.00%	0.10484	229.02%

5 Updates from QA@CLEF 2006

As 2006 was LCC's first year participating in CLEF, there were some substantial errors that were corrected for 2007 as well as some other improvements to various components of the system.

5.1 PowerAnswer Improvements

Answer type detection

We extended PowerAnswer's answer type detection module by moving it to a hybrid system which takes advantage of precise heuristics as well as machine learning algorithms for ambiguous questions. A maximum entropy model was trained to detect both answer type terms and answer types. The learner's features for answer type terms include part-of-speech, lemma, head information, parse path to WH-word, and named entity information. Answer type detection uses a variety of attributes such as additional answer type term features and set-to-set lexical chains derived from eXtended WordNet⁴ which links the set of question keywords to the set of potential answer type nodes.

Temporal processing

Dates for documents and the temporal context of the answer are maintained through question answering and after initial ranking, answers are given a boosting factor on top of their current relevance score that is intended to give greater priority to strong answers that are more recent than other strong answers. Answers that appear further down the response list and have lower relevance scores will not be affected by this boosting.

Because temporal answers can have a range of granularity, when pre-processing the data collection, the named entities stored in the IR index are extracted in a greedy fashion, so both "March 14, 1592" and "2000" will be tagged as *_date* to give PowerAnswer the best flexibility for entity selection. During answer processing, if the question is seeking just a month, or a year, then the excess information from the *_date* entity selected is removed after a more fine-grained NE recognition is performed on the answer nugget. EN $\rightarrow$ FR Q27 *In what year was Richard Nixon born?* demonstrates the utility of this method, where the answer is given in the text ... *naît le 9 janvier 1913* Otherwise, if a simple "When was ..." question is asked, the entity with the most detailed temporal information would be the final answer. This method operated on 4 EN $\rightarrow$ FR and 3 EN $\rightarrow$ PT questions seeking *year*, or *day*.

The temporal processing PowerAnswer performs [8] resulted in accuracy measures for temporally-restricted questions of 46.34% and 31.58% over French and Portuguese targets, respectively.

5.2 Machine Translation Improvements

Since last year (QA@CLEF 2006 evaluation), we improved the Answer Aligner module: (1) we fixed bugs that altered the order of the answer in the output and (2) we improved the alignment heuristics.

⁴ <http://xwn.hlt.utdallas.edu>

In terms of Machine Translation quality, we added modules in MeTRe designed to better preserve the structure of the sentence. The add-ons were focused on rules that can be easily derived from punctuation: numeric values, currency amounts, insertions through quotation marks and through brackets, etc.

5.3 Wikipedia Document Conversion

PowerHarvest is a tool developed by Language Computer Corp. that is used for document harvesting and preprocessing for Question Answering. One of the features of PowerHarvest is to convert XML database dumps⁵ into a format that is used by PowerAnswer's document collection indexing module.

Prior to QA@CLEF 07, PowerHarvest was limited to the English version of the Wikipedia collection – it only knew how to interpret English Wikipedia markup (e.g.: *Talk*, *User*, *User_talk*, *Template*, *Category*, ...). We extended PowerHarvest to work also on the targeted languages – French and Portuguese – by introducing support for French markup (e.g.: *Discuter*, *Utilisateur*, *Discussion_Utilisateur*, *Modèle*, *Catégorie*, ...) and Portuguese markup (e.g.: *Discussão*, *Usuário*, *Usuário_Discussão*, *Predefinição*, *Categoria*, ...).

The documents resulting from PowerHarvest (in French and in Portuguese) were translated using MeTRe and indexed, using the same procedures that were used for the Newswire parts of the collection (*Le Monde* and *French SDA* for French; *Público* and *Folha de São Paulo* for Portuguese – according to the *Guidelines for Participants in QA@CLEF 2007*).

6 Results

The integrated multilingual PowerAnswer system was tested on 200 English → French and 200 English → Portuguese factoid, list and definition 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 is the Confidence Weighted Score (CWS) that judges how well a system confidently returns correct answers.

Table 1 illustrates the final results of Language Computer's efforts in its participation at QA@CLEF for 2007.

7 Error Analysis and Challenges in 2007

While LCC saw a substantial improvement in errors over last year's results, there remain challenges that offer interesting research and engineering opportunities. The major sources of errors include: translation misalignments, tokenization errors, and data processing errors – questions and passages.

⁵ <http://download.wikimedia.org>

7.1 Translation Misalignments

Because the version of PowerAnswer used is *monolingual*, the system design for *multilingual* question answering involves translating documents dynamically for processing through the QA system and later mapping the responses back into the source language documents. This results 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, aggressive keyword expansion techniques diminish the impact of these mistranslations. Errors from misalignments still occurred due to

For the French source results, PowerAnswer returned 14 inexact answers, and for Portuguese source 7 inexact, 7% and 3.5% of the total response. Many of these inexact responses are definition-style questions that either

(1) did not have enough information, such as EN → FR Q158: *Who is Amira Casar?, actrice née le 1er juillet 1971 à Londres, d'une mère russe chanteuse d'opéra et d'un père d'origine kurde.* or (2) the alignment module was unable to correctly align the English answer within the given source language document, and so fell back to translating the English answer. While this particular default behavior is positive for the user since the answer is readable and still correct in nature, the language is not exact from the document and so warrants an inexact judgment in the evaluation. This failure is caused by translation errors when trying to map back from noisy text to the original source.

An example of this is EN → FR Q154: *Who is Allan Frederick Jacobsen?*. The source document is the Wikipedia “Allan Jacobsen” entry. The source language answer is *Allan Frederick Jacobsen, né le 22 septembre 1978 à Edimbourg (Écosse) est un joueur de rugby à XV qui joue avec l'équipe d'Écosse depuis 2002, évoluant au poste de pilier (1,78m et 109kg).*

The answer returned by PowerAnswer over the English translated Wikipedia article is *born on 22 September 1978 to Edinburgh (Scotland - is a player rugby to XV is playing with the team of Scotland since 2002 swimming as pillar (1.78 me and 109 kg).*

The final submitted result, which was translated as the default was *22 nés sur édimbourg à 1978 septembre un joueur - est (scotland est rugby xv à jouez avec écosse l ' équipe depuis 2002 de baigner (1.78 comme pilier 109 kg) moi et.* While the final answer is readable and comprehensible, it is not the answer as it appears in the source document.

7.2 Returning NIL as Answer

The version of PowerAnswer used for QA@CLEF uses parameters that relax some of the semantic and syntactic restrictions on answers that PowerAnswer uses when running on more stable and less noisy data. A result of this is that zero NIL answers were returned because the system always attempts to return an answer. An example of this is EN → PT Q13: *When did the blue whale become extinct?*, the answer to which is NIL because the blue whale has never become

extinct. PowerAnswer selected the translated answer *When the hunting of whale blue has finally been banned in the 1960s, 350000 whales Blue had been killed.* with the exact answer *the 1960s*, but with a low relative confidence score.

7.3 Other Error Sources

Other error sources are less specific to the methodology of intermediate translation and more general question answering errors such as answer type detection, keyword selection and expansion, passage retrieval and answer selection/ranking. An example of an answer selection error is EN $\rightarrow$ PT Q24 *What department is Caen the capital of?*. The correct answer string is *Caen é uma comuna francesa na região administrativa da Baixa-Normandia, no departamento Calvados* but PowerAnswer selected “Baixa-Normandia” as the correct answer instead of Calvados due to proximity.

7.4 English Accuracy

As we also included for last year’s results [1], Table 2 compares the PowerAnswer English factoid accuracy versus the mapped submission factoid accuracy. This table also demonstrates that the system did obtain the expected improvements after the correction of misalignment errors present in the submission for QA@CLEF 2006. Additionally, the list accuracy scores for this year were 30.00% (FR), 20.00% (PT); the definition scores were 22.22% (FR) and 25.81% (PT).

Table 2. LCC’s Factoid Results in English

Source	Submission Acc.	Eng. Position 1 Acc.
French	40.50%	52.06%
Portuguese	28.00%	39.23%

8 Conclusions

QA@CLEF 2007 proved to be a valuable learning exercise. We have been able to correct some of the errors that were present in last year’s results and achieve the kind of performance we expected from PowerAnswer operating on noisy translated data. Intermediate translation for question answering provides the opportunity for additional errors in processing, but we believe that our results in this evaluation show that such a methodology can be practical and accurate.

References

1. Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J., Moldovan, D.: LCC’s PowerAnswer at QA@CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 310–317. Springer, Heidelberg (2007)

2. Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A., Wang, P.: Employing Two Question Answering Systems in TREC-2005. In: Text REtrieval Conference (2005)
3. Koehn, P., Och, F.J., Marcu, D.: Statistical Phrase-Based Translation. In: Proceedings of HLT/NAACL 2003 Edmonton, Canada (2003)
4. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005 (2005)
5. 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)
6. Moldovan, D., Harabagiu, S., Clark, C., Bowden, M.: PowerAnswer 2: Experiments and Analysis over TREC 2004. In: Text REtrieval Conference (2004)
7. Moldovan, D., Clark, C., Bowden, M.: Lymba's PowerAnswer 4 in TREC 2007. In: Text REtrieval Conference (2007)
8. Moldovan, D., Bowden, M., Tatu, M.: A Temporally-Enhanced PowerAnswer in TREC 2006. In: Text Retrieval Conference (2006)
9. Moldovan, D., Clark, C., Harabagiu, S.: Temporal Context Representation and Reasoning. In: Proceedings of IJCAI, Edinburgh, Scotland (2005)
10. Moldovan, D., Clark, C., Harabagiu, S., Maiorano, S.: COGEX A Logic Prover for Question Answering. In: Proceedings of the HLT/NAACL (2003)
11. Moldovan, D., Novischi, A.: Lexical chains for Question Answering. In: Proceedings of COLING, Taipei, Taiwan (August 2002)
12. Moldovan, D., Rus, V.: Logic Form Transformation of WordNet and its Applicability to Question Answering. In: Proceedings of ACL, France (2001)
13. Moldovan, D., Srikanth, M., Fowler, A., Mohammed, A., Jean, E.: Synergist: Tools for Intelligence Analysis. In: NIMD Conference, Arlington, VA (2006)
14. Olteanu, M., Suriyentrakorn, P., Moldovan, D.: Language Models and Reranking for Machine Translation. In: NAACL 2006 Workshop On Statistical Machine Translation (2006)
15. Olteanu, M., Davis, C., Volosen, I., Moldovan, D.: Phramer - An Open Source Statistical Phrase-Based Translator. In: NAACL 2006 Workshop On Statistical Machine Translation (2006)
16. Tatu, M., Iles, B., Moldovan, D.: Automatic Answer Validation using COGEX. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)

RACAI's Question Answering System at QA@CLEF2007

Dan Tufiş, Dan Ştefănescu, Radu Ion, and Alexandru Ceaşu

Institute for Artificial Intelligence, Romanian Academy
13, "13 Septembrie", 050711, Bucharest 5, Romania
{tufis,danstef,radu,aceausu}@racai.ro

Abstract. We present a pattern-based question answering system for Romanian that participated in the Romanian monolingual task of the QA@CLEF 2007 track. We aim to prove that working with a good Boolean searching engine and using question type driven answer extraction heuristics and structural matching between the linkage representations of the question and candidate sentences, one can achieve acceptable results (30% overall accuracy).

Keywords: question answering, tokenization, tagging, lemmatization, indexing and retrieval, Lucene, answer extraction, query generation.

1 Introduction

The "attention economy", as the web services and content providing industry is also called, centered on the needs of a more and more demanding and busy user, has, as its central challenge the realization of ever better performing search and indexing engines, able to use semantic criteria for understanding of the retrieval requests and for finding the relevant documents. Natural language remains the most common way for the specification of an information request. Language technologies spectacularly evolved in the last two decades or so, but on the other hand, the volume of textual data which a practical NLP system is supposed to deal with, is several orders of magnitude larger and much noisier than the standard input data for the systems of the '80ies. The tension between the requirements for a deep NLP processing system and the response time expected by the average users of the electronic sources of information available on the web is still impossible to conciliate, in spite of remarkable increase of the computers' speed seen in the last years. It is not surprising that formalisms and their associated techniques which in the blooming era of Artificial Intelligence were considered to be completely inadequate for dealing with natural languages, nowadays are back in business: finite state technologies, statistical methods, shallow parsing techniques, pattern matching and so on. The user, sincerely interested in finding an answer to his query, uses, in the vast majority of cases, simple and direct questions, more often than not, below the level of the system' linguistic competence. This man-machine communication behaviour, referred to as "Pygmalion" effect, or computer-speak (by analogy to *foreign-speak* or to *baby-speak*) has been noticed long time ago [10]. Most of the current QA systems, including the state-of-the-art ones, are explicitly or implicitly based on the "computer-speak" paradigm and the work reported here is no exception.

The typical architecture of a QA engine, built on top of a search engine, additionally includes:

(i) a question analysis component that usually identifies the type of the question (factoid, definition, enumeration, etc.), the type of the expected answer (specifically for factoid questions: person, location, date, organization, etc.), decides on the terms to be used in query formulation and generates a query in the interrogation language of the search engine which the QA system is based on;

(ii) a paragraph extraction and ranking component; the documents returned by the search engine, that match the user's query, are processed to identify the sub-document textual units which could contain the expected answer; these textual units (usually paragraphs) are sorted according to the matching scores and the top N segments are further processed for answer extraction;

(iii) an answer extraction module which analyses the best ranked textual units in order to retrieve the complete, minimal and syntactically well-formed string(s) that presumably constitute(s) the answer(s) to the user's natural language query.

There are several problems to be solved for implementing a real-time open-domain QA system. The dichotomy speed-complexity in language processing has been successfully solved for some of these problems, such as part of speech tagging, lemmatization, named entity recognition or syntactic analysis. For others, mainly in the semantic area, better solutions are needed in order to achieve significant further progress. Current research in QA acknowledges these problems and it is no wonder that modern QA systems are extremely complex, comprising modules that deal with different levels of natural language representations. Systems such as FALCON ([3]), COGEX ([7]), LASSO ([8]), PowerAnswer ([2,6]), LaSIE ([1]) etc.) use some form of logical representation of both question and candidate answers in order to logically prove that the selected answers can be justified in terms of the question premises.

In this paper we will describe a pattern-based QA system, which is a combination of two distinct classifiers and answer extraction modules (A and B) built on top of the same text search engine. The two sub-systems have different strategies to build the queries for the search engines being specialized on different types of questions. As their performances are comparable and they do not make identical errors, combining them was both theoretically and practically motivated.

Our system participated in the QA@CLEF2007 evaluation exercise for the Romanian language obtaining the highest score on this language track.

2 The Document Collection, Indexing and Searching

The document collection was composed of 43486 Romanian language documents from Wikipedia. The files provided for the shared task were available both in HTML and XML formats (<http://ilps.science.uva.nl/WikiXML/>). The titles and contents from each document in the collection were pre-processed in order to obtain sentence and word splitting, part of speech tagging (POS tagging) and lemmatisation using the TTL (Tokenizing, Tagging and Lemmatizing) platform [4]. After the TTL run, we parsed the entire document collection using our link analyser LexPar [4,5]. This dependency-like analysis is called a linkage and it is produced with a link filter adaptation (LexPar) of

the Lexical Attraction Models (LAM) of Deniz Yuret ([12]). In principle, a LAM tries to assign to a given sentence the most likely undirected, acyclic, planar and connected graph where the vertices are the words of the sentence and its edges are the dependency links between words pairs.

The RACAI QA system uses a C# port of the Apache Lucene full-text searching engine. Lucene is a Java-based open source (under the Apache License) toolkit for text indexing and searching developed by Apache Jakarta project..

Although the Lucene toolkit comes with several already-made tokenizers, stemmers and stop word filters, we preferred to develop a custom indexing scheme using our own annotated resources. There were considerable improvements when we used the Romanian tokenizer instead of Lucene's default tokenizer because the compounds (words with hyphen) and the abbreviations were handled in a consistent manner. Also, relying on our language specific lemmatizer (instead of Lucene's stemmer) proved to be a source of significant improvement of the overall performance of the QA system. The TTL lemmatizer uses the POS-tagging information, because, in the vast majority of the cases (in Romanian), the part of speech solves the lemmatization ambiguities. As index terms we used only the words tagged as nouns, main verbs, adjectives, adverbs or numerals.

Each word in a document of the collection is indexed for both its occurrence form and the respective lemma as well as for their position (within the title and/or the document's body). These distinctions result in four different index fields: title word form (title), title lemma (ltitle), document word form (text) and document lemma (ltext).

Given a Boolean query with several conjunctive clauses, the system will first try to match all of the query clauses against the document index. If the search fails, the system will recursively try to match $n - 1$ of the conjunctive clauses until the query returns at least one result from the document index. The returned documents are used to select the corresponding sections in which the query terms occur.

3 The Sub-system A

As already mentioned, the first processing step in answering a question is processing the user's input and generating a formal query, intelligible by the text search engine. The more accurate this step, the better the chances that the search engine would return relevant snippets, out of which the answer extraction module might produce the correct answer. The sub-system A, in a more traditional way, uses a bag of words approach and takes into account the content words in the questions, the noun phrases formed by them and all the subparts of the noun phrases that start with a content word. All these are searched in lemma and word forms both in the title and the text, the query being obtained by concatenating them using the logical operator AND. As mentioned, the search engine was programmed to return the snippets that contain the majority of the terms.

The queries usually are enriched with synonyms extracted from a large lexical ontology (Romanian WordNet). Our system allows the user to specify whether or not the query should be synonymy expanded but, we found that this did not improve the results. One possible explanation for this rather surprising observation was that most of the questions were formulated using the same words as in the expected answers.

One important aspect of the standard way of processing open-domain questions is the detection of the type of expected answer. The precise identification of the type of the expected answer facilitates a more accurate extraction of the answer from the snippets returned by the search engine. To this end, the sub-system A uses a Maximum Entropy classifier [9] which given a set of features extracted from the current question computes the most likely class of the expected answer. We took into account features like: the first WH word (*cine* - who, *unde* - where, *când* - when, *care* - which, *ce* - what, *cum* - how, *cât* - how many), the existence of other words before the WH word, of certain verbs at the start of the sentence (like *numi* - name), the existence of a word denoting measurement units, the existence of a word denoting temporal units, the existence of the verb “to be” as the first verb, the existence of at least two non-auxiliary verbs, the existence of a proper noun as the first noun or not, the number of the first noun, the part of speech of the first content word (noun, verb or numeral), the punctuation mark at the end of the question if different from question mark. Our classifier considered 8 types of expected answers: temporal (TMP), time interval (ITMP), definition (DEF), measure (MES), list (LST), location (LOC), names (N) and explanation (WHY). We manually classified 500 questions according to these classes and used them for training. The classifier was tested and fine-tuned on the training data. At run time, for the shared task evaluation, the module correctly labeled 199 questions out of 200 with respect to the type of the expected answer. The pattern-based approach used for answer extraction worked very well for definition questions while for the other types the results were much poorer. Therefore, in the overall combined QA system the sub-system A was credited for answering DEF-type questions while the other types were subject to combination with the results of the second sub-system. The answer extraction module embedded into the sub-system A was implemented as a pattern matcher ruled by a set of patterns dependent on the type of the expected answer. For the DEF questions in Romanian, we noticed that, usually, the focus of the question is the first NP found in the question starting with a common noun, a proper noun, or an adjective and therefore, the first NP with those properties was automatically set as the focus of the question. The word form or the lemma form of the focus was looked up in every sentence of the sections of the documents returned by Lucene and we looked for several positive or negative clues as (i) the existence of “to be” verb (along with a possible auxiliary) immediately following the focus and the existence or not of indefinite articles or demonstrative pronouns or articles after the verb, (ii) the existence of an opened left bracket immediately after the focus, followed or not by a noun, (iii) the existence of a comma in front or after the focus, (iv) the existence of certain prepositions before the focus or (v) the existence of a definite oblique article in front of the focus. Since the definitions are usually found in the beginning of the documents we penalize the candidates as we find them farther and farther in the document. When the focus was found only in lemma form or only partial matching was found, the candidates were again penalized. The rank of a candidate, computed based on the number of query terms matched by the candidate, was another selection criterion. Altogether, these criteria provided a weighted base scoring function. Different combinations of the positive clues led to the weighting of the total score with values between 0.6 and 1.5.

4 The Sub-system B

The second sub-system adopted a very different approach, compensating for the brute-force and ad-hoc solutions as implemented in the sub-system A.

Similarly to the sub-system A, the input question is preprocessed by TTL to obtain word tokenization, part of speech tagging and lemmatization. Unlike the first sub-system, sub-system B does not use a bag-of-words approach, but instead it uses Lex-Par to generate a dependency linkage of the question. The linkage is used to extract the focus-topic articulation of the question. This operation is based on grammatical patterns, defined as sequences of POSes labeling the words of the question forming a dependency chain. Here are the most important patterns for Romanian (considered from the beginning of the question):

- 1 {prep}, {WH determiner}, {noun(FOCUS)}, {main-verb}, {noun(TOPIC)}
- 2 {WH pronoun(FOCUS)}, {main verb}, {noun(TOPIC)};
- 3 {WH adverb(FOCUS)}, {main verb}, {noun(TOPIC)};
- 4 {main verb}, {noun(FOCUS)};

For instance the reading of the first pattern is: *in a dependency chain formed by an (optional) preposition, linked to a WH-type determiner, linked to a noun which is linked to a main verb, further linked to a noun, the first noun is the FOCUS and the second noun is the TOPIC.*

After the focus and topic are extracted, the query for Lucene text search engine is created by following the links in the linkage of the question in order to extract all the links that are formed between content words (nouns, main verbs, adjectives and adverbs). With this list of links at hand, the query is computed as a logical disjunction of terms in which each term corresponds to a content word to content word link and it is equal to a logical conjunction of the lemmas at the end points of the link.

Consider the question “În/prep ce/wh-det localitate/noun s-/refl-pron a/v-aux născut/v-part Leonardo/noun da/prep Vinci/noun?” for which the linkage of the content words is {<localitate, născut>, <născut, Leonardo>, <Leonardo, Vinci>} thus resulting in the following query: **(Ittext:localitate AND Ittext:naște) OR (Ittext:naște AND Ittext:Leonardo) OR (Ittext:Leonardo AND Ittext:Vinci).**

Answer extraction is basically the best structural match between the linkage of the question and the linkage of each of the sentences in the paragraphs that have been returned by the text search engine. As the linkage is not a full dependency parse, we arbitrarily choose the first main verb in the sentence/question to be the root of the linkage. To better explain the structural matching between the linkage of the question and the linkage of one document sentence let us follow the example in Figure 1.

For the question “În ce localitate s-a născut Leonardo da Vinci?”¹ the text search engine returns among other paragraphs, one which begins with the sentence “**Leonardo** s-a **născut** la 14 aprilie 1452 , nu departe de Florența, în mica **localitate** Anchiano.”² (the keywords from the question are bolded). In Figure 1, on the left side we have the linkage of the question (functional words removed) and on the right, we have the linkage of the candidate sentence (functional words also removed). Structural match means going depth-first through the question tree one node at the time and for each such node (let it be Nq), going depth-first through the sentence tree searching

¹“What town was Leonardo da Vinci born in?”

²“Leonardo was born on April 14 1452, not far from Florence, in the little hamlet of Anchiano.”

from the current node in the question tree (let N_s be the matching node). When such a node is found, a matching score S is increased by $1/(1 + |\text{depth}(N_q) - \text{depth}(N_s)|)$ such that if the nodes are at the same depth in the two trees, the value of S increases by 1. Otherwise, the value of S increases by the inverse absolute difference of depths at which matching nodes are found. For the two trees in Figure 1, $S = 3$ (see the dotted arrows which mark 3 matching nodes at the same depths).

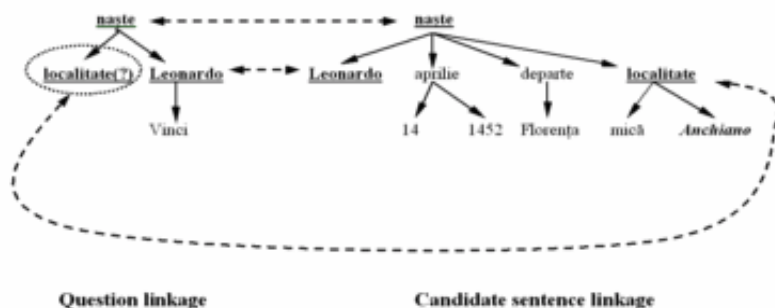


Fig. 1. Structural match between the question “În ce localitate s-a născut Leonardo da Vinci” and one candidate sentence

After the structural match score S is computed, we extract all the subtrees from the candidate sentence tree such that: a) subtrees do not contain already matched nodes (with the only exception being the subtree with the focus node as root if it exists) and b) they are at the same depth as the focus node of the question (marked with the “?” in Figure 1). For our example, we have three such subtrees: “14 aprilie 1452”, “(nu) departe (de) Florența” and “(în) mica **localitate** Anchiano”. If we have the subtree that is governed by the focus, the answer is given by the corresponding named entity in this subtree which in our case is the town “Anchiano” (a proper noun). In the opposite case (we do not have a subtree with the focus as root), we extract all entities consistent with the POS of this topic-focus articulation (here, a proper noun) and propose all of them as possible answers.

Structural match occurs between the linkage of the question the linkage of each sentence of each paragraph that was returned by the text searching engine. We want to order the candidate sentences by the S score but also by the score of the paragraph in which the sentence occurred (we want to also give credit to the text searching engine). This way, the final candidate sentence score is $A = \alpha S + (1 - \alpha)P$ where P is the score of the paragraph containing the candidate sentence (with the best results obtained for α to 0.4) Answers are thus extracted (as explained above) from the top candidate sentences ordered by the A score.

5 Results and Conclusions

The evaluation of our QA system revealed two major sources of further improvement.

The first one refers to the query formation and the results returned by the LUCENE search engine. It is obvious that if none of the snippets returned by the search engine

contains the relevant information, the answer extraction phase could not but fail. The search engine worked significantly harder for the sub-system A than in case of sub-system B. For the answer extraction, both sub-systems considered the top 10 snippets as ranked by LUCENE search engine. For the sub-system A, the query formation resulted in 50 cases (out of 200) of non-relevant returned snippets while for the sub-system B there were 46 cases. This strongly suggests that although much shorter, the queries generated based on dependency linking managed to convey in a much precise way the semantics of the questions. However, we were surprised to notice that the bag-of-words approach returned the most relevant snippet in the first position in 136 cases while for the dependency-linking approach only 128 queries had the same result. One reason we found was that the dependency linking, failed on some occasions to set the right links between the significant words of the question. Another reason was that in some cases, by eliminating the intervening functional words, not all the content words were connected (that should have been linked), and the generated queries missed one or more significant search terms. And finally, the subsystem B, unlike the sub-system A, did not take into account the titles, which definitely represent a strong relevancy clue for important search terms. One way of improving these deficiencies (besides analyzing the titles) would be to train the dependency linker on much larger corpora and also to apply some kind of link "transitivity" when intervening functional words are removed.

The second source of improvement refers to the indexing criteria. The LUCENE indexes were created having in mind a bag-of-words approach, not a dependency based model of sentence analysis. As such the indexing was more appropriate for the sub-system A than for sub-system B. One line of our future investigations would be focused on texts indexing based on dependency links/relations.

We should mention that some of the questions to be answered were related. These questions were arranged into groups, for most of the questions, to retrieve relevant results we had to take into account information found in the previous questions. We managed to handle this situation by adding the query generated for a new question to the query of the first question of the group.

Table 1. The official results (R=right, W=wrong, X=inexact, U= unsupported)

	First Run	Second Run
Questions	Total:200; 60 R; 105 W; 34 X; 1 U Overall accuracy = 60/200 = 30.00%	- 60 R;101 W; 39 X; 0 U Overall accuracy = 60/200 = 30.00%
Factoids	Total: 160; 38 R; 90 W; 31 X; 1 U Accuracy = 38/160 = 23.75%	Total: 160; 38 R; 86 W; 36 X; 0 U Accuracy = 38/160 = 23.75%
Lists	Total: 10; 0 R; 10 W; 0 X; 0 U Accuracy = 0/10 = 0.00%	Total: 10; 0 R; 10 W; 0 X; 0 U Accuracy = 0/10 = 0.00%
Definition Questions	Total: 30; 22 R; 5 W; 3 X; 0 U Accuracy = 22/30 = 73.33%	Total: 30; 22 R; 5 W; 3 X; 0 U Accuracy = 22/30 = 73.33%
Temporally Restricted Questions	Total: 51; 10 R; 31 W; 10 X; 0 U Accuracy = 10/51 = 19.61%	Total: 51; 10 R; 31 W; 10 X; 0 U Accuracy = 10/51 = 19.61%
NIL Answers Returned	Total: 54; 7 R; 47 W; 0 X; 0 U Accuracy = 7/54 = 12.96%	Total: 54; 7 R; 47 W; 0 X; 0 U Accuracy = 7/54 = 12.96%

For the Romanian QA track we submitted two runs of the system with two sets of parameters and weights for snippets ranking and answer extraction. The difference in behaviour was negligible. In accordance with the official evaluation, which took into account only the first answer (see Table 1), the best answered questions were those of type definition. From the total of 200 questions, 31 were identified as requiring a DEF type answer but only 30 were in fact of this nature. Our QA system answered correctly the DEF questions in 25 cases although in 3 situations the answers were considered inexact (possibly because of their length). For another 4 questions one could have found the correct answer among the first three solutions.

We should notice that we did not answer correctly to any of the *Lists* questions because, at the time of the evaluation, the answer extraction module lacked the ability to coalesce relevant segments from different sentences.

Acknowledgements

The work reported here was developed within the ROTEL and SIR-RESDEC projects granted by the National Authority for Scientific Research.

References

1. Gaizauskas, R., Humphreys, K.: A Combined IR/NLP Approach to Question Answering Against Large Text Collections. In: 6th Content-Based Multimedia Information Access Conference (RIAO 2000), Paris, France, pp. 1288–1304 (2000)
2. Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A., Wang, P.: Employing Two Question Answering Systems in TREC 2005. In: Text Retrieval Conference (TREC-14), Gaithersburg, Maryland (2005)
3. Harabagiu, S., Moldovan, D., Paşca, M., Mihalcea, R., Surdeanu, M., Bunesco, R., Gîrju, R., Rus, V., Morărescu, P.: FALCON: Boosting Knowledge for Answer Engines. In: Text Retrieval Conference (TREC-9), Gaithersburg, Maryland, pp. 479–489 (2000)
4. Ion, R.: Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis, Romanian Academy, Bucharest (2007)
5. Ion, R., Barbu Mititelu, V.: Constrained Lexical Attraction Models. In: Nineteenth International Florida Artificial Intelligence Research Society Conference, pp. 297–302. AAAI Press, Menlo Park (2006)
6. Moldovan, D., Bowden, M., Tatu, M.: A Temporally-Enhanced PowerAnswer in TREC 2006. In: Text Retrieval Conference (TREC-15), Gaithersburg, Maryland (2006)
7. Moldovan, D.I., Clark, C., Harabagiu, S.M., Hodges, D.: COGEX: A semantically and contextually enriched logic prover for question answering. *J. Applied Logic* 5(1), 49–69 (2007)
8. Moldovan, D., Harabagiu, S., Paşca, M., Mihalcea, R., Goodrum, R., Gîrju, R., Rus, V.: Lasso: A Tool for Surfing the Answer Net. In: Text Retrieval Conference (TREC-8), Gaithersburg, Maryland, pp. 175–184 (1999)
9. Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution. PhD thesis, University of Pennsylvania, Philadelphia, PA (1998)
10. Slator, B.M., Anderson, M.P., Conley, W.: Pygmalion at the interface. *Communications of the ACM* 29(7), 599–604 (1986)
11. Voorhees, E.M.: Overview of the TREC 2005 Question Answering Track. In: Text Retrieval Conference (TREC-14), Gaithersburg, Maryland (2005)
12. Yuret, D.: Discovery of linguistic relations using lexical attraction. PhD thesis. MIT, Cambridge (1998)

DFKI-LT at QA@CLEF 2007*

Bogdan Sacaleanu, Günter Neumann, and Christian Spurk

German Research Center for Artificial Intelligence (DFKI)

Language Technology Lab

Saarbrücken, Germany

{bogdan,neumann,cspurk}@dfki.de

Abstract. In this paper we present our QA@CLEF 2007 version of QUANTICO, a cross-language open domain question answering (QA) system for German and English document collections. The main features of the system are: use of a preemptive off-line document annotation with linguistic information; online extraction of abbreviation-expansion pairs and appositional constructions for the answer extraction; use of online translation services for the crosslingual scenarios; use of redundancy as an indicator of good answer candidates; selection of the best answers based on distance metrics defined over graph representations of the answer's context. The results of evaluating the system's performance by QA@CLEF 2007 were as follows: for the German-German run we achieved an overall accuracy (ACC) of 30 %; for English-German 18.5 % (ACC); for German-English 7 % (ACC), for Spanish-English 10 % (ACC) and for the Portuguese-German run 7 % (ACC).

1 Introduction

QUANTICO is a cross-language, open domain question answering (QA) system for factoid and definition questions. The system was originally developed for German and English crosslingual and multilingual tasks in a uniform framework; at this year's QA@CLEF competition we have evaluated the system on Spanish and Portuguese as further source languages for the first time. The language barrier in the multilingual scenarios is crossed on the question side rather than on the document side through the use of free online translation services, alignment and other linguistic resources. An offline preprocessing adds several layers of linguistic information to the target document collection in advance: named entity and sentence boundary annotations may thus be easily and efficiently used in the later answer document retrieval process leading to more accurate and more reliable answer document collections. The answer extraction is based on these retrieved document collections; final answer candidates are extracted from these

* The work presented in this paper was partly supported by the European Commission, contract FP6-33860, Question Answering Learning technologies in a multiLingual and Multimodal Environment (QALL-ME, <http://qallme.fbk.eu/>) and by a research grant from the German Federal Ministry of Education and Research (BMBF) to the DFKI project HyLaP (<http://hylap.dfki.de/>, FKZ: 01 IW F02).

collections using redundancy as the principal indicator for suitability. After a normalization of the answer candidates, a selection component chooses the best answer by representing the context of each answer candidate as a graph and computing the answer's appropriateness in terms of the distance between the answer and the question's keywords (Sacaleanu and Neumann 2006).

A Wikipedia snapshot was added to the CLEF answer document collection this year which we preprocessed in the same way as done before for the news articles corpora. For the integration of the two new source languages Spanish and Portuguese we pursued another strategy than we did in previous evaluations. So far we always analyzed the source questions first and used alignment methods then to cross the language barrier. For the new source languages, however, we opted for first translating the questions and then interpreting them.

In the next section we will begin with a very brief overview of the QA system's architecture and the processing of factoid and definition questions in both monolingual as well as crosslingual scenarios. Section 3 will then go into more detail by introducing the principal system components and their internals. In Sect. 4 we present the results of the QA@CLEF evaluation campaign and conclude the paper with a quick analysis of the remaining issues with the system's performance.

2 System Overview

The QUANTICO system provides a uniform framework for monolingual and crosslingual QA scenarios. This section briefly describes the workflow of the system for the QA@CLEF 2007 track before Sect. 3 details the components that are used in this workflow.

A novelty in QA@CLEF 2007 were topic question clusters that often contain anaphoric references between the questions. To deal with such topic questions in QUANTICO we first annotate these questions with named entities (NEs) and link personal pronouns to the corresponding NEs.¹ Therewith we can treat all questions in the same way, no matter whether they belong to a topic cluster or not.

In the crosslingual scenarios each question is then translated into the target language. Therefore we use three freely available online translation services (AltaVista Babelfish², FreeTranslation.com³ and Traduction Voila⁴) for each question. The possible translations are then interpreted independently of each other in a question analysis step. The results of these analyses are ranked according to linguistic well-formedness and their completeness with respect to inquiry information (question type, question focus and expected answer type); the best of the three ranked analyses is used for further processing. There is a notable difference in the workflow for the different source target language pairs, though: for certain language pairs there are no translation services available in which case

¹ For both NE annotation and pronominal coreference resolution we have used LingPipe 1.7 (cf. <http://www.alias-i.com/lingpipe/>).

² <http://babelfish.altavista.com/>

³ <http://www.freetranslation.com/>

⁴ <http://trans.voila.fr/>

we use English as an interlingua. For our QA@CLEF 2007 participation this means that we had to translate the Portuguese source questions into English first before eventually translating them into German in a second step.

In the monolingual scenarios each question is directly processed by our question analysis component. Thus, no matter whether we are in a crosslingual or in a monolingual scenario, we always end up with a question analysis in the target language. The next step is to use this formal representation of the question to retrieve potential answer documents and extract answer candidates from these documents based on their number of occurrence. For factoid questions, where the answers are usually NEs or simple chunks, the answer extraction is slightly different than the answer extraction for definition questions, where answers may range from simple chunks to whole sentences.

Finally, the best answer candidate is selected as the returned answer according to distance metrics that are computed using features of the question’s keywords and the answer candidates.

3 Component Descriptions

This section is a description of QUANTICO’s individual components that have been used in this year’s evaluation exercise along with some examples. The component and module names that are used in the following subsections refer to the system architecture diagram in Fig. 1.

3.1 NE-Informed Translation

Named entities (NEs) can cause problems in translations by being translated when they actually should not be translated. To achieve more accurate translation results, the translation component of QUANTICO includes a substitution module that replaces some NE types with placeholders before actually translating the question; after the translation, the replacement is reversed. The drawback of this approach is that the outcome of the substitution module and thus the overall translation highly depends on the accuracy of the NE recognizer, since an inaccurate markup of the NE terms may prevent from translating semantically relevant information.

As an example for the NE-informed translation consider the NE-annotated question in (1) that shall be automatically translated to German: each NE is replaced by a placeholder for which it is unlikely that it is translated by a machine translation service – see sentence (2). After the automatic translation to German, (3), all placeholders can be replaced again with the original NEs as in (4).

- (1) When did <person>White</person> become CEO of
 <organization>Wiley and Sons</organization>?
- (2) When did Smith become CEO of ACME?
- (3) Wann wurde Smith Geschäftsführer von ACME?
- (4) Wann wurde <person>White</person> Geschäftsführer von
 <organization>Wiley and Sons</organization>?

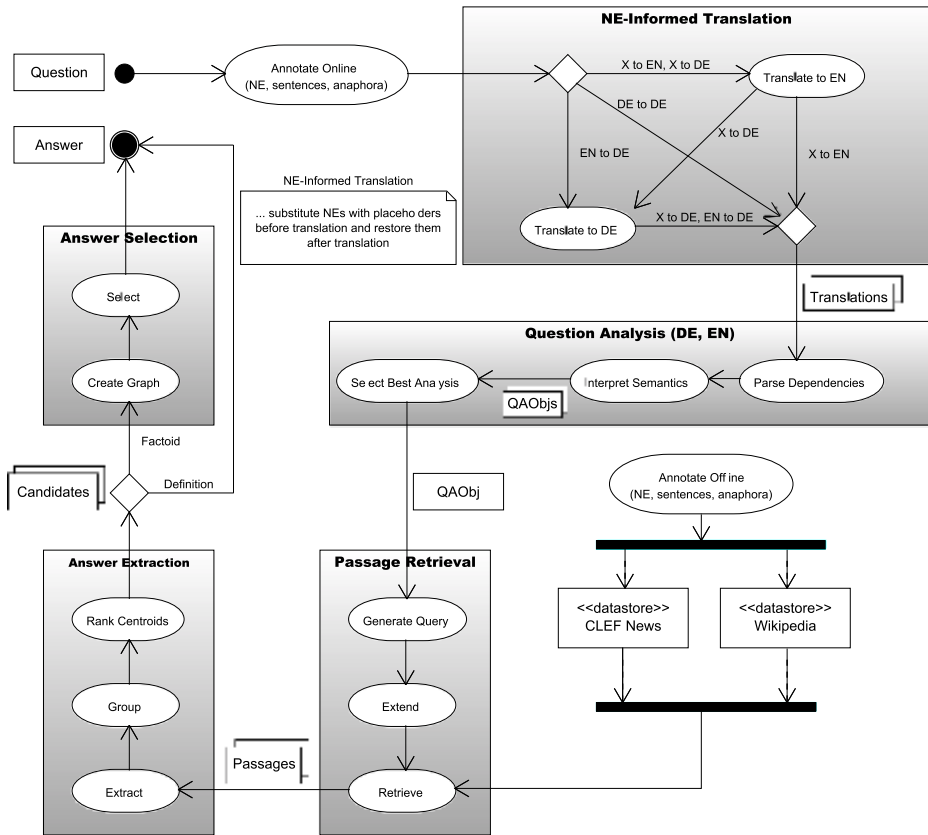


Fig. 1. System Architecture

In the case of unavailable online translation services for certain pairs of languages, such as for Portuguese-German, an interlingua solution has been implemented: English is then used as an intermediate translation, i.e., a Portuguese question is first translated into English and then from English into German.

3.2 Question Analysis

The question analysis component has basically two principal functions: analyzing the incoming question syntactically and semantically. The *Parse Dependencies* process performs the syntax analysis and returns a syntactic dependency tree which also contains recognized named entities (NEs). In the *Interpret Semantics* and the *Select Best Analysis* processes, a semantic question analysis object (QAObj) is computed which contains information like question type, expected answer type and question focus (cf. Neumann and Sacaleanu 2005 for details).

The semantic information is determined on the basis of syntactic constraints applied to relevant parts of the dependency tree (e.g., considering agreement and functional roles) and by taking into account information from two small knowledge bases (Neumann and Piskorski 2002). The latter basically perform a mapping from linguistic entities to semantic question analysis information; this may be trigger phrases like `name_of`, `type_of`, `abbreviation_of` or mappings from lexical elements to expected answer types like `president` $\rightarrow$ `PERSON`. For German, we additionally perform a *soft retrieval match* to the knowledge bases by performing an online compound analysis combined with string similarity tests. For example, from the lexical mapping `Stadt` $\rightarrow$ `LOCATION` for the lexeme “Stadt” (town), we automatically derive mappings from the nominal compounds “Hauptstadt” (capital) and “Großstadt” (large city) to `LOCATION`, too.

3.3 Passage Retrieval

The *Annotate Offline* process is a prerequisite for the passage retrieval component. It annotates the document collections preemptively with information that may be valuable during the *Retrieve* process. Since for example the expected answer type for factoid questions is usually a named entity (NE) type, annotating the documents with NEs provides for an additional indexation unit that may help to narrow down the number and range of retrieved passages⁵ only to those documents that contain the required answer type.

The *Generate Query* process mediates between the question analysis result (QAObj) and the search engine which serves the answer extraction component with information units (Passages). The *Generate Query* process creates different kinds of IR queries depending on the question type. Each IR query makes use of the advanced indexation units in its own way. As an example, consider the question in (5 a); since NEs were annotated during the offline annotation and were then used as indexation units, the query generator now builds an IR query which restricts the search only to those passages that have at least two `LOCATION` NEs as can be seen in (5 b): one for the possible answer (“Berlin”) and the other for the question’s keyword “Germany”.

- (5) a. What is the capital of Germany?
 b. `+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 they share no lexical overlap. For a question like “Who is the French prime minister?” there may be answering passages like “prime minister X of France”, “prime minister X [...] the Frenchman” and “X, the French government leader”. In order to find such relevant passages with lexically different question keywords, the *Extend* process was developed. It accounts

⁵ Information units or “passages” in our CLEF 2007 system were always single sentences.

for bridging the lexical gap through look-up of related words: for the former example, looking at the keyword “French” it might extend the IR query with the words “France”, “Frenchman” and “Frenchwoman”.⁶

3.4 Answer Extraction

The answer extraction component is based on the assumption that the redundancy of information is a good indicator for an answer’s suitability. For factoid and definition questions the component extracts different kinds of answers: simple chunks (i.e., NEs and basic NPs) for factoid questions and complexer structures (simple phrases up to whole sentences) for definition questions. Based on the QA control information from the QAObj, different extraction strategies are triggered in the *Extract* process: factoid question usually trigger NE extraction while definition questions trigger an extraction of those passages that resemble a definition. The extraction of such potential definition answer passages is attained by matching them against a lexico-syntactic pattern of the form

<searched concept> <definition verb> .+

where **<definition verb>** is a verb coming from a closed list of verbs like “be”, “mean”, “signify”, “stand for” etc.

Besides the plain extraction of answer candidates in the answer extraction component, a ranking of the candidates is performed; this is done in two steps: the answer candidates are clustered in a first step and then these clusters are ranked in a second step. The first step is carried out in the *Group* process where different mentions of the same semantic answer are clustered. For factoid questions, where the candidates are usually NEs or chunks, the computation 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, the clustering consists in finding out the focus of the explanatory sentence or the head of the considered phrase.

In the *Rank Centroids* process each cluster eventually gets a weight 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 selects a list of top answers based on a distance metric defined over the answer’s context. This context is first normalized by removing all functional words and is then represented as a graph

⁶ For our CLEF 2007 system we have only extended the IR queries with country name related words as in the example. For each English and German we have used lists of words corresponding to about 200 different countries.

structure with the tokens at the nodes (*Create Graph* process).⁷ 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 graph; after calculating this score in the *Select* process, the answers are sorted by this score and the best-scored answer is finally chosen and returned.

4 Evaluation Results and Error Analysis

At QA@CLEF 2007 we have participated in five tasks for each of which we have submitted one run only: DEDE (German to German), ENDE (English to German), DEEN (German to English), ESEN (Spanish to English) and PTDE (Portuguese to German). A summary of the achieved results can be found in Table 1.

Table 1. System Performance for QUANTICO at QA@CLEF 2007

Run ID	Right		Wrong	Inexact	Unsupported
	#	%	#	#	#
dfki071dede _c	60	30.0	121	14	5
dfki071ende _c	37	18.5	144	18	1
dfki071deenc _c	14	7.0	178	6	2
dfki071esen _c	10	5.0	180	10	0
dfki071ptdec _c	5	2.5	189	4	2

Several known issues of the QUANTICO system that have already been uncovered in previous evaluations could not be addressed timely for the QA@CLEF 2007 track. So the error analysis presented in Sacaleanu and Neumann (2006) is still largely up to date. Further issues that we have found in a preliminary error analysis of the 2007 results are the following:

- The English dependency parser has a coverage which is smaller than initially assumed and therefore created a bottleneck for those runs with English as the target language.
- The *NE-Informed Translation* component which is highly dependent on the accuracy of the named entity (NE) recognizer has led to very bad results for the runs with Spanish and Portuguese as the source language; for these runs the German NE recognizer was used by default. In these cases a simple translation, i.e., without replacing any NEs, would certainly have been more successful.

⁷ For CLEF 2007 we only used a simple “textual” graph here with all token nodes being connected to the nodes of their adjacent tokens from the sentence context. The possibility of using a dependency graph was dropped for the competition for performance reasons.

References

- Sacaleanu, B., Neumann, G.: DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track. In: Working Notes for the CLEF 2006 Workshop, Alicante, Spain, September 20–22 (2006)
- Neumann, G., Piskorski, J.: A Shallow Text Processing Core Engine. *Journal of Computational Intelligence* 18(3), 451–476 (2002)
- Neumann, G., Sacaleanu, B.: Experiments on Robust NL Question Interpretation and Multilayered Document Annotation for a Cross-Language Question/Answering System. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 411–422. Springer, Heidelberg (2005)

University of Wolverhampton at CLEF 2007*

Georgiana Puşcaşu and Constantin Orăsan

Research Group in Computational Linguistics
University of Wolverhampton, UK
{georgie,c.orasan}@wlv.ac.uk

Abstract. This paper reports on the participation of the University of Wolverhampton in the Multiple Language Question Answering (QA@CLEF) track of the CLEF 2007 campaign. We approached the Romanian to English cross-lingual task with a Question Answering (QA) system that processes a question in the source language (i.e. Romanian), translates the identified keywords into the target language (i.e. English), and finally searches for answers in the English document collection. We submitted one run of our system that has achieved an overall accuracy of 14%, and a precision over non-NIL answers of 33.73%. Error analysis revealed that this low performance is mainly due to the lack of a reliable translation methodology from the source in the target language.

1 Introduction

Cross-lingual Question Answering is defined as the task of retrieving the answer in one language (the target language) to a question posed in a different language (the source language). Last year, a new Romanian-to-English (RO-EN) cross-lingual QA task was organised for the first time within the context of the CLEF campaign [10], and it consisted of retrieving answers to Romanian questions from a collection of English documents. This year's task [6] was similarly organised, with the exception that all questions were clustered in classes related to the same topic, some of which even contain anaphoric references to other questions from the same topic class, or to their answers. Besides the usual news collections employed in the search for answers, this year's novelty was the fact that Wikipedia articles could also be used as answer source.

This is the first time a Romanian-English cross-lingual QA system fully developed at the University of Wolverhampton has participated in the QA@CLEF competition. This system contains the classical QA modules: question processing, information retrieval and answer extraction [7]. In addition, the cross-lingual capabilities are provided by a Romanian-to-English term translation module. This paper describes the development stages and evaluation results of our system. The rest of the paper is organised as follows: Section 2 provides an overall description of the system, while Sections 3, 4, 5 and 6 present the four embedded modules - the question processor, term translator, passage extractor and answer extractor respectively. Section 7 captures the evaluation results and their analysis. Finally, in Section 8, conclusions are drawn and future directions of system development are considered.

* This work has been supported by the EU funded project QALL-ME (FP6 IST-033860).

2 System Overview

Question Answering systems normally share a pipeline architecture consisting of three main stages: question analysis, passage retrieval and answer extraction [7]. For cross-lingual systems, the language barrier is usually crossed by employing free online translation services for translating the question from the source language into the target language [8,14]. The QA process is then entirely performed in the target language by a monolingual QA system. A different approach taken by some cross-lingual systems automatically translates the document collection in the source language, performs monolingual QA in the source language [2], and then converts the answer back into the target language by aligning it with the corresponding span in the original document. Another alternative approach involves monolingual QA in the source language and then translating the answer, but this approach is feasible only when document collections covering the same material are available in both the source and target languages [1].

Since we could not identify reliable translation services from Romanian into English for translating complete questions, nor English-Romanian full document translation tools, the first two approaches were discarded. For the third option, the impediment was the lack of a Romanian document collection equivalent to the English one. Therefore we adopted a slightly different methodology where the question analysis is performed in the original source language without any translation in order to overcome the negative effect of full question translation on the overall accuracy of the system. Afterwards, in order to link the two languages involved in the cross-lingual QA setting, term translation is performed by means of bilingual resources and linguistic rules. The search for passages and answers is then performed in the target language documents. This method was also employed by Sutcliffe et al. [13] and Tanev et al. [14].

The system architecture consists of a four-module pipeline, where each module is responsible for a different stage in answering a question. These four modules are:

- 1) ***Question Processor***

This module analyses each Romanian question in order to identify the type of the question and of the expected answer, the question focus, and all relevant keywords.

- 2) ***Term Translator***

For each question term, all translation equivalents are generated by consulting bilingual resources and by employing linguistic rules to assemble individual words into target language terms.

- 3) ***Passage Extractor***

At this stage candidate snippets of text are retrieved from the English document collection on the basis of a query that includes the translation equivalents of all terms identified in the question.

- 4) ***Answer Extractor***

On the basis of the information extracted by the Question Processor, this module identifies in the previously retrieved snippets a set of candidate answers matching the expected answer type. One answer is then selected by ranking the resulting set of candidate answers.

The following four sections present in more detail the functionality of each module.

3 Question Processor

This module is mainly concerned with the identification of the semantic type of the entity sought by the question, but it also provides the question type, focus, and relevant keywords. To achieve these goals, our question processor performs the following steps:

a) Question Annotation: The questions are first morpho-syntactically pre-processed using the TnT POS tagger [3] trained on Romanian [16], and afterwards noun phrases (NPs) and named entities (NEs) are identified using a rule-based approach. Temporal expressions (TEs) are also detected using the adaptation for Romanian of an English TE identifier and normalizer [11].

b) Question Focus Identification: The question focus is considered to be either the noun determined by the question stem 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 (EAT): Our system can detect the following expected answer types: PERSON, LOCATION, ORGANIZATION, TEMPORAL, NUMERIC, DEFINITION and GENERIC. The assignment of a class to an analysed question is performed using the question stem and the question focus type. The latter is obtained using Romanian WordNet [17] sub-hierarchies specific to the categories PERSON / LOCATION / ORGANIZATION.

d) Inferring the Question Type: This year, the QA@CLEF main task distinguishes among four question types: *factoid*, *definition*, *list* and *temporally restricted* questions [6]. As temporal restrictions can constrain any question type, we first detect whether the question has the type *factoid*, *definition* or *list*, and then search for temporal restrictions. The question type is identified as follows: for questions which ask for definitions of concepts, the assigned question type is *definition*; if the question focus is a plural noun, then the question type is *list*, otherwise it is *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 generated by listing the question terms in decreasing order of their relevance, as follows: the question focus, the identified NEs and TEs, the remaining NPs, and the non-auxiliary verbs. This relevance ranking is not currently employed at the retrieval stage, but it will be used in the future to assign weights to each term. Given the grouping of questions into topics and the presence of anaphoric expressions between same topic questions, a shallow anaphora resolution mechanism is employed to expand the keyword set with other possibly relevant terms as described below. The expanded set of keywords is then passed on to the term translation module, in order to obtain English keywords for passage retrieval.

f) Resolution of anaphoric expressions: As related questions are organised in clusters, in a number of cases, the links between questions are realised using anaphoric pronouns, and therefore, in order to obtain a more complete list of keywords, anaphora resolution is necessary. Given the difficulty of the task, it is not possible to employ a fully fledged anaphora resolution system. Instead, the set of keywords related to a question is expanded with the list of NEs present in the cluster. This is done for two reasons. On the one hand, investigation of the question clusters revealed that pronouns often refer to NEs in the cluster. On the other hand, given that the questions are related, it is possible

that NEs present in the questions also co-occur in the same document. As a result, it is more likely to extract relevant documents with this expanded query. Certain questions referred to the answer of the previous question. Currently, this problem is not addressed because in our present system there is no way to feed an answer back into the system.

4 Term Translator

Each keyword is translated into several translation equivalents, which are then grouped using the disjunction operator into a keyword specific sub-query. The conjunction of all sub-queries corresponding to the question keywords forms the final query.

Term translation is achieved with an approach similar to the one we employed when we participated together with two Romanian research groups in the same task at CLEF 2006 [12]. It also resembles the one employed by Ferrandez et al. [5] for the English to Spanish task of the same CLEF campaign. This method employs WordNet and the ILI alignment between the English WordNet and the other WordNets developed in the EuroWordNet and BalkaNet projects. The underlying idea is that, given a Romanian word, the Romanian WordNet and its alignment to the English one, we identify all possible translations of the word by finding all the synsets it appears in and extracting the equivalent English synsets through the ILI alignment. If the word to be translated does not appear in the Romanian WordNet, as is quite frequently the case, we search for it in other dictionaries and preserve the first three translations. If still no translation is found, the word itself is considered as translation, an approach which works reasonably well for NEs. In the case of multi-word terms, each word is translated individually using the method described above. After that, rules are employed to convert the Romanian syntax into English syntax, and to obtain the translation equivalents of a given term.

One drawback of this method is that, by not employing word sense disambiguation, it proposes too many translations for a word. To address this problem, we implemented a ranking method which relies on parallel English-Romanian Wikipedia pages and on the assumption that the two sets of pages will contain more or less the same information, so it will be possible to find the most likely translation for a given term. Unfortunately, preliminary experiments revealed that, by including this approach, a very small number of passages are retrieved, many of which do not contain the answer to the question. Due to time restrictions, we were unable to properly tune the method to retrieve better passages, and for this reason we did not employ it in this year's submission.

5 Passage Extractor

The purpose of this module is to extract a list of passages which may contain the answer to a given question from the following three document collections: English Wikipedia pages collected in November 2006, Los Angeles Times from 1994 and Glasgow Herald from 1995. This is the first time that Wikipedia has been included in the document collection and, as a result of the fact that it is several orders of magnitude bigger than the other two collections, the search space was significantly larger than in previous years, making the task more difficult. Given that the documents in each collection are formatted in different ways, each had to be indexed individually and processed in a

slightly different manner. For indexing and retrieval, we used Lucene [9], an open source information retrieval library.

Passages are extracted using the query proposed by the term translation module, including all possible translations of the question keywords. In the initial experiments we limited the number of translations used for each original keyword, but as a result, the number of retrieved snippets was too low. This can be explained by the fact that no disambiguation was performed and therefore it was possible that some of the translations were ranked high and included in the query, even though they were not appropriate. As the attempt to order the translations according to their likeliness of being the correct translation of a keyword did not lead to satisfactory results, it is not used in this year's submission. In light of this, we decided to consider all the translations identified for a keyword and link them with the OR operator provided by Lucene.

We indexed the collection in order to retrieve documents containing the keywords, and not actual passages. This approach is taken because it offers more flexibility and allows better control of the methods which retrieve candidate passages. It has the drawback that it needs to process each document individually and extract relevant passages. For this year's system only sentences are extracted. In order to do this, each sentence from the retrieved documents is scored on the basis of how many keywords, TEs and NEs it contains. At present, up to 25 sentences with the highest scores are retrieved from each document, provided that their score is higher than a predefined threshold. This set of sentences is fed into the next module, the answer extractor.

6 Answer Extractor

Once candidate answer-bearing document passages have been selected, the answer extractor starts by merging all passages retrieved for questions belonging to a certain topic. All retrieved passages are parsed with Conexor's FDG Parser [15] and with the NE identifier embedded in the GATE toolkit [4]. A question-based passage ranking is then applied to the merged set of passages to identify the most relevant passages. The answer extractor then addresses each EAT in a different manner, as follows:

a) *Expected answer type is a Named Entity*: Named entities having the desired answer type are identified in the retrieved passages and added to the set of candidate answers. Candidate answers are then ranked on the basis of the passage score, the distance to other keywords and their frequency. The candidate answer with the highest score is presented as final answer. When the retrieved passages contain no candidate answer, the system returns NIL.

b) *Expected answer type is NUMERIC*: Several NUMERIC answer sub-categories are distinguished: MONEY, PERCENTAGE, MEASURE and NUMERIC-QUANTITY (any other NUMERIC entity). Patterns are defined for exact candidate answer identification, patterns that take into consideration either the format of certain numeric expressions or the presence of the question focus in the neighbourhood of a numeric expression. The process of ranking candidate answers relies on the same parameters as in the case of the Named Entity answer type.

c) Expected answer type is TEMPORAL (i.e. a Temporal Expression): The subtypes of TEMPORAL entities that guide the answer extraction process are: MILLENNIUM, CENTURY, DECADE, YEAR, MONTH, DATE, TIME, DURATION (applying also to questions asking about age) and FREQUENCY. If the granularity of the expected answer is coarser than the granularity of a candidate answer TE, patterns are employed to convert the TE to the required granularity (e.g. if the EAT is YEAR and the candidate answer has the granularity DATE like “25th of January 1993”, then only “1993” is extracted).

d) Expected answer type is GENERIC: When the EAT is neither a NE, nor a NUMERIC or TEMPORAL entity, the question focus is essential in finding the answer. The candidate answers are constrained to be hyponyms of the question focus head.

e) Expected answer type is DEFINITION: A different approach is taken when the question asks for the definition of a concept. Wikipedia contains definitions for a large number of concepts, therefore our first attempt is to obtain the definition from the Wikipedia page corresponding to that concept. To this end, Lucene is used to return Wikipedia pages which contain in their title words from the concept to be defined. Because this approach returns more than one document, a ranking method is applied to the retrieved documents. The more concept words the document title contains, the more the document score gets boosted. Once the documents are ranked, patterns are used to locate the answer. Whenever no answer can be located in Wikipedia, passages are extracted from the other two document collections using the passage extractor described in Section 5 and the regular expressions are then applied to them. Unfortunately, this fall-back approach performed quite poorly.

7 Evaluation Results

This section describes the results corresponding to the run we submitted for the RO-EN QA task at CLEF-2007. The methodology employed targets precision at the cost of recall, by providing NIL answers to those questions we cannot reliably locate a candidate answer in the retrieved passages. Apart from this, no more than one answer per question is returned, and this is the first ranked answer, when it can be identified.

Table 1 illustrates the detailed results achieved by our system. Despite the fact that our Question Processor is able to recognise questions asking for LISTS, the answer extractor does not tackle this type of questions. The overall accuracy of our system was evaluated at a generic score over all questions of 14%. An analysis of the system output revealed the fact that our system was unable to locate an answer and returned the answer NIL for 117 questions. It retrieved 83 answers, out of which 28 correct, 49 wrong, 4 unsupported and 2 inexact.

A preliminary analysis of the incorrect and NIL answers showed that their main cause was the poor translation of the question keywords, this yielding either irrelevant or no passages being retrieved from the English document collection. If we consider the fact that our system, whenever it has little or no confidence that it has found a correct answer, does not attempt to answer the question by returning NIL, and we analyse only the answers retrieved by the system, the conclusion is that out of 83 answers, 28 are correct, this yielding a precision of 33.73%.

Table 1. Detailed evaluation results

	FACTOID	LIST	DEFINITION	TEMPORALLY RESTRICTED
RIGHT	15	0	13	0
WRONG	140	9	17	2
UNSUPPORTED	4	0	0	1
INEXACT	2	0	0	0
TOTAL	161	9	30	3
ACCURACY	9.32%	0.00%	43.33%	0.00%

Unsupported answers are correct answers, but the returned support passage is not considered relevant enough for the question. Given that we can not access the correct answers and expected support passages, it is difficult to judge whether the four retrieved passages are appropriate or not. For example, in the case of the question “*What kind of animal did Victor Bernal try to buy on the 25th of January 1993?*”, our returned answer was “*gorilla*”, and it was extracted from: “*The sting took place on Jan. 25, 1993, when Bernal and the others were escorted onto a DC-3 cargo plane parked in a remote corner of a small Miami airport to see the gorilla, crated for shipment.*”, which seems correct, but probably it can not be justified only by the presence of Bernal’s name, of the date mentioned in the question, and of the noun “*gorilla*”, which is a type of “*animal*”.

In the case of inexact answers, the answer-string contains the correct answer and the provided snippet supports it, but the answer-string is incomplete or more detailed than the correct answer. For example, given the question “*What is the occupation of Michael Barrymore?*”, our inexact answer was “*troubled comic*” and the supporting passage was “*Troubled comic Michael Barrymore last night received an ovation as his show, Strike It Lucky, was named Quiz Programme of the Year at the National Television Awards.*”. Most of these errors can be corrected by improving the answer extractor with more specific rules as to the extent of the required answer.

8 Conclusions

This paper describes the development stages of our cross-lingual Romanian to English QA system that participated in the QA@CLEF campaign. Adhering to the generic QA architecture, our system implements the three essential stages (question processing, passage retrieval and answer extraction), as well as a term translator which provides cross-lingual capabilities by translating question terms from Romanian into English. This year our emphasis was less on fine tuning the system, and more on exploring the issues posed by the task and developing a complete system able to participate in the competition. Therefore, all four modules are still in a preliminary stage of development.

The run we submitted for the Romanian to English cross-lingual QA task achieved an overall accuracy of 14%, the best score achieved among systems with English as target language [6]. An in-depth analysis of the results at different stages in the QA process has revealed a number of future system improvement directions. The term translation module has a crucial influence over the systems performance, and will therefore receive most of our attention. Apart from this, we will further investigate the ranking method for translation equivalents which relies on information from parallel

English-Romanian Wikipedia pages in order to improve its performance, as we believe it is a promising research direction. We also intend to improve our answer extraction module by identifying a better answer ranking strategy.

References

1. Bos, J., Nissim, M.: Cross-Lingual Question Answering by Answer Translation. In: Working Notes for the CLEF 2006 Workshop (2006)
2. Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J., Moldovan, D.: LCC's PowerAnswer at QA@CLEF 2006. In: Working Notes for the CLEF 2006 Workshop (2006)
3. Brants, T.: TnT - a statistical part-of-speech tagger. In: Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP 2000), Seattle, WA (2000)
4. 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)
5. Ferrandez, S., Lopez-Moreno, P., Roger, S., Ferrandez, A., Peral, J., Alvarado, X., Noguera, E., Llopis, F.: AliQAn and BRILI QA Systems at CLEF 2006. In: Working Notes for the CLEF 2006 Workshop (2006)
6. Giampiccolo, D., Forner, P., Penas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: Working Notes for the CLEF 2007 Workshop (2007)
7. Harabagiu, S., Moldovan, D.: Question Answering. In: Mitkov, R. (ed.) Oxford Handbook of Computational Linguistics, pp. 560–582. Oxford University Press, Oxford (2003)
8. Jijkoun, V., Mishne, G., de Rijke, M., Schlobach, S., Ahn, D., Muller, K.: The University of Amsterdam at QA@CLEF 2004. In: Working Notes for the CLEF 2004 Workshop (2004)
9. LUCENE, <http://lucene.apache.org/java/docs/>
10. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Osenova, P., Peas, A., Jijkoun, V., Sacaleanu, B., Rocha, P., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Working Notes for the CLEF 2006 Workshop (2006)
11. Puscasu, G.: A Framework for Temporal Resolution. In: Proceedings of the 4th Conference on Language Resources and Evaluation (LREC 2004) (2004)
12. Puscasu, G., Iftene, A., Pistol, I., Trandabat, D., Tufis, D., Ceausu, A., Stefanescu, D., Ion, R., Orasan, C., Dornescu, I., Moruz, A., Cristea, D.: Cross-Lingual Romanian to English Question Answering at CLEF 2006. In: Working Notes for the CLEF 2006 Workshop (2006)
13. Sutcliffe, R., Mulcahy, M., Gabbay, I., O'Gorman, A., White, K., Slattery, D.: Cross-Language French-English Question Answering using the DLT System at CLEF 2005. In: Working Notes for the CLEF 2005 Workshop (2005)
14. Tanev, H., Kouylekov, M., Magnini, B., Negri, M., Simov, K.I.: Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005. In: Working Notes for the CLEF 2005 Workshop (2005)
15. Tapanainen, P., Jaervinen, T.: A Non-Projective Dependency Parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, ACL (1997)
16. Tufis, D.: Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In: Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 1105–1112 (2000)
17. Tufis, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Tufis, D., ed.: Romanian Journal on Information Science and Technology. Special Issue on BalkaNet. Romanian Academy (2004)

Bilingual Question Answering Using CINDI_QA at QA@CLEF 2007

Chedid Haddad and Bipin C. Desai

Department of Computer Science and Software Engineering
Concordia University
1455 De Maisonneuve Blvd. W.
Montreal, Quebec H3G 1M8, Canada
{c_haddad,bcdesai}@cs.concordia.ca

Abstract. This article presents the first participation of the CINDI group in the Multiple Language Question Answering Cross Language Evaluation Forum (QA@CLEF). We participated in a track using French as source language and English as target language. CINDI_QA first uses an online translation tool to convert the French input question into an English sentence. Second, a Natural Language Parser extracts keywords such as verbs, nouns, adjectives and capitalized entities from the query. Third, synonyms of those keywords are generated thanks to a Lexical Reference module. Fourth, our integrated Searching and Indexing component localises the candidate answers from the QA@CLEF data collection. Finally, the candidates are matched against our existing set of templates to decide on the best answer to return to the user. Out of eight runs submitted this year, CINDI_QA ranked second and third with an overall accuracy of 13%.

Keywords: Question answering, Questions beyond factoids, Bilingual, French, English.

1 Introduction

The Concordia Index for Navigation and Discovery on the Internet (CINDI [1]) group has been founded at the Department of Computer Science and Software Engineering of Concordia in the late 1990s. Its purpose is the continuous enhancement of information discovery and retrieval. Hence, CINDI_QA has been created to tackle bilingual question answering within the spectrum of QA@CLEF 2007.

CINDI_QA is composed of one main logical entity, the Processor, which is plugged to several existing tools that help it understand and analyze the input question. Additionally, templates are highly relied upon to put together the best possible answer and return it to the user.

The paper is organized in the following way: we first go through the system overview with an emphasis on architecture. We then list the tools incorporated in CINDI_QA. Afterwards, we take a look at the template matching mechanism that drives the system and its application for the QA@CLEF participation and the results obtained. The conclusion follows where we highlight what we learned and how we intend to improve CINDI_QA's performance for the future editions of QA@CLEF.

2 System Overview

CINDI_QA is made up of one central unit called the Processor, a couple of peripheral components, four integrated tools and a template module. These are illustrated in figure 1 and elaborated on in the next section.

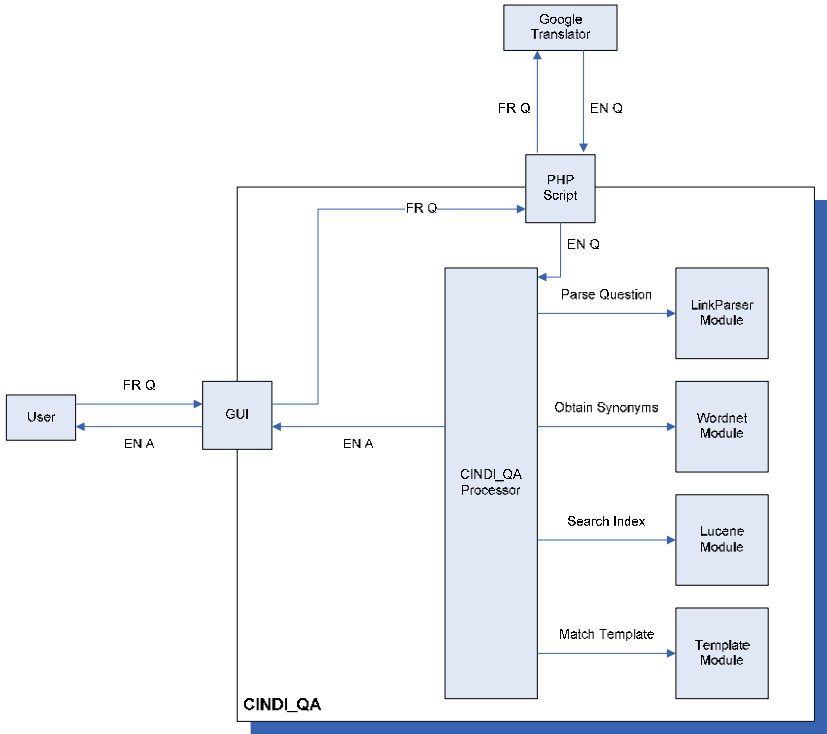


Fig. 1. CINDI_QA Architecture

The system's "brain" is located in the CINDI_QA Processor where all the logical inferences are made. Since CINDI_QA communicates with several tools, the information it retrieves from those modules - which is meaningless on its own - is analyzed in the Processor in a structured way that helps build the eventual answer. This process is done in a pre-defined order, with the Processor getting data from one module, sorting it out then using it to probe the next module.

One of the peripheral components is the PHP Script that acts as an interface between the Online Translator and the Processor. Its purpose is to send the French question to the Online Translator and bring its English equivalent back. The other peripheral unit is the Graphical User Interface (GUI) which is the façade of our system for the user; it will prompt for a French question and return the English answer.

In a typical scenario, the question introduced in French by the user is translated to English then parsed to extract the keywords. Afterwards, the synonyms of the keywords are obtained to produce an internal query sent to the Search engine that already has the CLEF data indexed. The candidate answers are localized at which point they are matched against a pre-existing set of templates that enables the selection of the best answer. That answer, which is in English, is sent back to the user.

In order to improve performance, CINDI_QA allows user interaction to direct its flow of operations. Actually, CINDI_QA can be run in two modes. In the automatic mode, it acts as a complete black box by only taking a question and returning an answer. In the feedback mode, it has the possibility of disambiguating the query by prompting the user after each stage. For example, since certain synonyms of a word often do not suit a particular context, the user can choose to eliminate them and only retain relevant ones.

The process flow of CINDI_QA in the feedback mode is shown in figure 2.

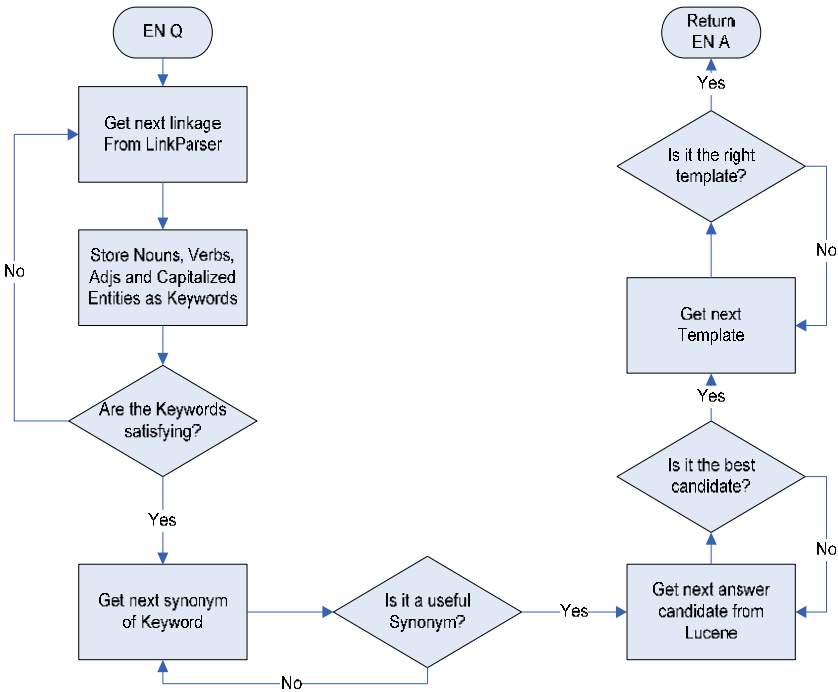


Fig. 2. CINDI_QA Process Flow

3 Tools Integration

3.1 Online Translation

Since we are working in a bilingual environment, the system is queried in a language different from the data collection it is using as reference. A translation tool is needed

for this reason. After researching the available tools, we noticed that the Google [2], Babel Fish [3] and Systran [4] translators appear to be powered by the same engine because they offer the same result when asked to translate from French to English.

We chose to use Google Translate in our system due to its better interface and speed of processing. A PHP script is responsible of delivering the French question typed in by the user to the Google Translate webpage and bring back the translated English equivalent to the CINDI_QA Processor.

3.2 Natural Language Parsing

We need a way to understand the question asked by the user in order to single out the keywords. This was achieved thanks to the Link Grammar Parser [5], a syntactic parser of English based on link grammar, an original theory of English syntax. Given a sentence, the Link Parser assigns to it a syntactic structure which consists of a set of labeled links connecting pairs of words. Each structure is called a linkage and several linkages are generated for the same input. This tool is written in generic C code, but since CINDI_QA has been programmed in Java, we used the Java Native Code Link Grammar Interface [6].

The Link Parser is plugged into our system to generate linkages of the English question. Using one linkage, we are able to determine which words form nouns, verbs, adjectives and capitalized entities. If those keywords appear wrong or incomplete, we go on to the next linkage. Remember that the user has the option of choosing the most appropriate linkage.

3.3 Lexical Reference

To increase the chances of finding candidate answers among the data collection, we include synonyms of the keywords in addition to the keywords themselves. WordNet [7] is a lexical database for the English language developed at Princeton University and has been used in other CINDI related projects [8] so its selection was pretty obvious.

WordNet was used in concordance with Lucene. Lucene enables us to create an index composed strictly of synonyms defined by WordNet that can be queried like a regular index [9], so we can actually get a list of synonyms of a specific word.

After defining the keywords using the Link Parser, we queried the WordNet index to obtain the synonyms of each keyword, except for capitalized entities. Since some of those synonyms are irrelevant or out of context, the user has the choice to discard them and select only the appropriate ones.

3.4 Document Indexing and Searching

We needed a tool that could not only index all the documents that make up QA@CLEF's data collection but also search the created index with some level of intelligence such as ranking results and highlighting query terms. A perfect match for this requirement is the Apache Software Foundation's Lucene [10], a high-performance, full-featured text search engine library written entirely in Java.

CINDI_QA makes extensive use of Lucene. As mentioned before, it is used in concordance with WordNet to get synonyms of keywords. Lucene also creates the CLEF index and ranks the results found. The following features are of great importance to our system.

3.4.1 Lucene Query Building Using Proximity Search

Once we have identified the keywords and their synonyms, the building of the query takes place. The query is constructed by putting together each keyword or its synonym with the other keywords or their synonyms. The crucial point here is to add the proximity search flag to the built query so that Lucene will not look for a sentence that has our keywords adjacent to each other, but rather one where the keywords are close to each other but spread out in a paragraph. This is done by adding the tilde character '~' and a number to the end of the query.

3.4.2 Highlighting Query Terms in Candidate Answers

Once the Lucene query is built, it is searched against the index of the document collection. Lucene then returns a list of filenames ranked according to the frequency of occurrence of the words in the query. At this point, we take advantage of the Lucene Highlighter [9], a wonderful tool that actually displays snippets of text with the query terms highlighted. This allows us not only to know which document has the answer, but also to obtain a sentence in that document that displays the actual answer.

4 Template Matching

CINDI_QA's template module comes into play in the final stages, between the identification of the candidate answers and the return of an answer to the user. Indeed, after the Lucene component's job is done, we are left with a set of sentences one of which holds the actual answer. To determine which one to choose, we match them against our set of pre-defined templates. We chose to use templates because a previous project [8] done by a member of our group was pretty successful at parsing English questions using mainly templates.

According to the CLEF Working Notes [11], there are three different types of questions: Factoid, Definition and List. Factoid and Definition kinds are further broken down into sub-categories. Our templates take advantage of this approach to differentiate among the questions that are inserted in our system, and also of the keywords we identified in the Link Parser module. The capitalized entity, made up of one or more words whose first letter is uppercase, identifies proper nouns. It is an important keyword because of its high frequency of occurrence; this is due to the fact that most questions at QA@CLEF ask about famous people, locations or events.

4.1 The Person Definition Template: *Who Is X?*

This is without a doubt the easiest template to consider. To discover that a question is of that type, it must start with the word "who", have the verb "is" or its derivative "was" and must be followed by a capitalized entity X. X could be one word or a composition of words; as long as adjacent words are capitalized, CINDI_QA will define them as belonging to one and the same capitalized entity. The following illustrates the query given to CINDI_QA, the candidate returned by the Highlighter of the Lucene Module and the answer returned after template matching.

- Example:
- query: *Who is Robert Altman?*
 - candidate: Robert Bernard Altman (February 20, 1925 – November 20, 2006) was an American film director known for making films that are highly naturalistic, but with a stylized perspective.
 - answer: American film director.

4.2 The Count Factoid Template: *How Many N V?*

This template is selected if the question starts with the words “how many” and has at least one verb V and one noun N, capitalized entities being optional. In the following example, had the expression “Solar System” not been capitalized, the Link Parser module would have identified “solar” as an adjective and “system” as a noun, yet the template would still be selected because “planets” is a noun.

- Example:
- query: *How many planets does the Solar System have?*
 - candidate: The Solar System or solar system consists of the Sun and the other celestial objects gravitationally bound to it: the eight planets, their 165 known moons, three currently identified dwarf planets (Ceres, Eris, and Pluto) and their four known moons, and billions of small bodies.
 - answer: Eight.

4.3 The Time Factoid Template: *What Year V1 X V2? / When V1 X V2?*

This template is selected if the following occurs: the question starts with the words “what year” or “when”, then has two verbs, V1 and V2 as well as one capitalized entity X. The Link Parser module will identify two separate verbs even if they belong to the participate form of the same verb, i.e. it will flag “was” as V1 and “murdered” as V2, even though technically they both define the verb “murder”. In the following example, notice how the word “assassinate”, synonym of the introduced verb “murder”, is used to return the correct answer.

- Example:
- query: *What year was Martin Luther King murdered?*
 - candidate: On April 4, 1968, King was assassinated in Memphis, Tennessee.
 - answer: 1968.

In the feedback mode, the user will be prompted one last time to approve the template chosen by CINDI_QA before being shown an answer. In the black-box mode, the system chooses the least costly linkage from the Link Parser, retains only the first two synonyms of a keyword, uses the best ranked document from Lucene with the most highlighted sentence and constructs the answer based on the first template matched.

5 Results and Analysis

The CINDI group participated in the FR to EN track of the QA@CLEF edition of 2007. We produced two runs that were sent: cind071fren and cind072fren. Both runs were similar, only differing in the length and detail level of the answer string; hence they ended up with the same overall accuracy value of 13%. Table 1 specifies the different assessments both runs obtained.

Table 1. CINDI_QA Results at QA@CLEF 2007

	Right	Wrong	Inexact	Unsupported	Unassessed	Accuracy
cind071fren	26	171	1	2	0	13%
cind072fren	26	170	2	2	0	13%

In addition to the usual news collections, articles from Wikipedia were also considered as answer source for the first time this year. But we became aware of that very late in our answering process and were only able to use a small part of the corpora available. This is the main reason for our system’s low performance.

Since our two runs are equivalent, table 2 shows the accuracy by question type only of cind072fren.

Table 2. CINDI_QA Results by Question Type

	Factoids	Lists	Definitions
Total	161	9	30
Right	18	1	7
Wrong	140	8	22
Unsupported	2	0	0
Inexact	1	0	1
Accuracy	11.18%	11.11%	23.33%

CINDI_QA ranked second and third out of eight participating runs this year [11]. The first ranked system has only a 1% better overall accuracy, but is better than us in Definition questions by 20%.

Definition questions are the most easy to answer because they are short and not very complex, hence the relatively high score of all systems to answer this type of questions. Having missed a large part of the corpora and because Definition questions weren’t abundant, our score in that category is average compared to our peers.

However, when it comes to Factoid questions, CINDI_QA holds the top spot with 11.18%. This is mainly due to the fact that most of our templates simulate Factoid questions so our system can handle them more efficiently.

6 Conclusion and Future Works

Since this is our first participation in QA@CLEF and the CINDI group only invested 1 man/year on the project, we have high hope for the future, especially since we missed a large part of the source corpora.

We learned that because CINDI_QA relies on so many external tools, it is only as strong as its weakest link. For instance, if from the start, the translation of the question is not a successful one, there is nothing the system can do after that stage to come up with a correct answer.

Future works include the addition of new templates to handle the multitude of question sub-categories as well as a mechanism to identify questions whose answer is not located in the CLEF data collection and return NIL for those questions.

References

1. The CINDI System, http://cindi.encs.concordia.ca/about_cindi.html
2. Google Translate, http://translate.google.com/translate_t
3. Babel Fish Translation, <http://babelfish.altavista.com/>
4. Systran Box, <http://www.systransoft.com/>
5. The Link Grammar Parser, <http://www.link.cs.cmu.edu/link>
6. Java Native Code Link Grammar Interface, <http://chrisjordan.ca/projects>
7. WordNet, a lexical database for the English language, <http://wordnet.princeton.edu/>
8. Stratica, N.: NLPQC: A Natural Language Processor for Querying CINDI, Master Thesis, Concordia University (2002)
9. Gospodnetic, O., Hatcher, E.: Lucene in Action. Manning, Greenwich (2005)
10. Lucene, <http://lucene.apache.org/>
11. Giampiccolo, D., Peñas, A., Ayache, C., Cristea, D., Forner, P., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, P., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: Cross Lingual Evaluation Forum Working Notes (2007)

The University of Évora's Participation in QA@CLEF-2007

José Saias and Paulo Quaresma

Departamento de Informática
Universidade de Évora, Portugal
{jsaias,pq}@di.uevora.pt

Abstract. The University of Évora participation in QA@CLEF-2007 was based on the Senso question answer system. This system uses an ontology with semantic information to support some operations. The full text collection is indexed and for each question a search is performed for documents that may have one answer. There is an ad-hoc module and a logic-programming based module that look for answers. The solution with the highest weight is then returned. The results indicate that the system is more suitable for the definition question type.

1 Introduction

This paper describes the use of SENSO Question Answer System in the Portuguese monolingual Question Answering (QA) task of this year's edition of Cross Language Evaluation Forum (CLEF). After two previous participations in 2004 [1] and 2005 [2], the Informatics Department of the University of Évora developed and tested this new system, based on the authors' previous work [3] and [4].

Besides the usual newspapers collections from Público and Folha de São Paulo, the system had to consider also the Portuguese articles from Wikipedia. It uses an ontology as a knowledge base with semantic information usefull in several steps along the process.

The next section explains the system architecture. The methodology is described with examples in section 3. The evaluation of the obtained results is presented in section 4. Finally, some conclusions and future work are pointed out in section 5.

2 System Architecture

Senso Question Answer System has five major modules: *Libs*, *Query*, *Solver*, *Ontology* and *Web Interface*. Figure 1 represents the way they are connected.

The *Libs Module* contains collections of text documents. These collections (Público and Folha de São Paulo from years 1994 and 1995, plus the Wikipedia documents) are seen as libraries that contain information needed for question

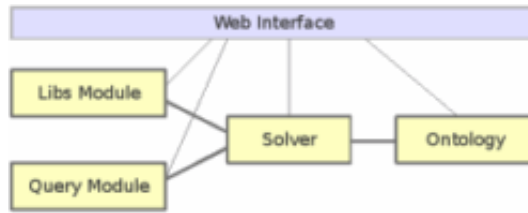


Fig. 1. Senso Modules

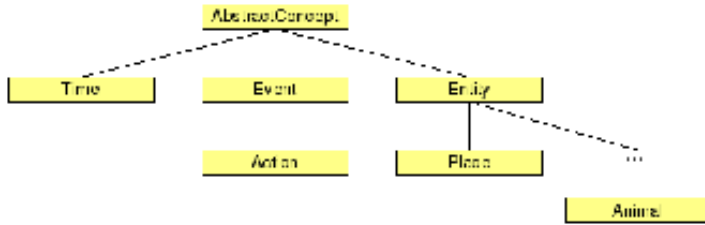


Fig. 2. Senso Ontology: top-level concepts

answering. All questions are firstly analyzed by the *Query Module*, which will select a set of relevant documents for each question, as explained later in section 3.

When we have an isolated sentence it's usually difficult to automatically capture its meaning. The *Senso Ontology* module has a starting knowledge base with semantic information that helps to perform the sentence analysis and the subsequent inference processes. This information is structured by an OWL¹ Ontology including concepts, relations and properties. Besides concepts and “IsA” relations, the ontology includes some simple facts about everyday life that might be very useful for text analysis. Our current ontology contains about 3500 concepts and has several relations connecting them: isA, usedFor, locatedAt, capableOf and madeOf. These concepts and relations represent a small common sense knowledge base about places, entities and events. Some of the top-level concepts are shown in figure 2.

The *Solver Module* performs a search for plausible answers in the identified relevant documents, being aware of the semantic expressed in the ontology. It has a logic-programming based tool and an ad-hoc answer selector.

The *Web Interface* layer allows an easier and friendly usage of the system, simplifying the analysis of each intermediate step in the process, as illustrated in figure 3. This interface is used to browse the ontology and make small changes to it, or to search for documents (or queries) and read them. Next section explains the methodology used to find the answers.

¹ OWL is the short name for Web Ontology Language. It is a language proposed by W3C to be used on *Semantic Web* for representation of ontologies.



Fig. 3. Web Interface: options for intermediate analysis

3 Methodology

This section explains our approach to the Question Answer track in detail.

3.1 Import the Text Collections

The starting point is the information source: the document collections, having more than 500000 texts. The XML collection files were processed and split in single texts, along with their metadata. The *Libs Module* keeps all these individual documents, being aware of their temporal context, which is obtained from the collection.

Because we needed to perform some text search operations, the collections were indexed at this point with Lucene², a full-featured text search engine library. Each text was processed with PALAVRAS[5], a syntactical parser³ based on the Constraint Grammars formalism that has a good coverage of the Portuguese language. This tool gives a detailed morpho-syntactical representation of the text for later usage.

3.2 Question Analysis

Each question is processed with the syntactical parser PALAVRAS[5] and a semantic analyzer able to obtain a partial semantic representation. The technique used for this process is based on Discourse Representation Structures (DRS) [7]. The partial semantic representation of a sentence is a DRS built with two lists,

² Apache Lucene is an open source project. <http://lucene.apache.org/>

³ Tool developed by Eckhard Bick. VISL Project: <http://visl.hum.sdu.dk/visl>

Let us consider the following definition question, in this year's edition:

Figure 4 shows the morpho-syntactical representation given for that question. We can see the parser tags identifying the subject, the predicate and the interrogative form *quem* (*Who*). Figure 5 has the DRS for the same question, with the semantic representation used by the system for later logic inference process. That means the system will search for someone whose name is *Boaventura Kloppenburg*.

The *Query Module* creates the Lucene search query. This is done with the question text terms and, for some, their related terms. So, if a question has something like “*Which bird...*” the text search query will include synonyms of *bird* and specialization terms given by the Senso ontology, such as *eagle*. This semantic operation in the query allows the retrieval of a text that may not have the word *bird* but is still relevant as a possible answer source. As an example, one question asked which tree is present in the Lebanon flag. The answer was *cedro* (or cedar, in English). Being aware that cedar is a tree was important to the process.

Fig. 4. Syntactical Parser: sample output

Fig. 5. DRS for Question ‘Who is Boaventura Kloppenburg?’

When the question belongs to a cluster and it is not the first from that group, the query is fed with more terms, in order to include the implicit topic. The system goes back to that cluster’s first question and gets their search terms and answer into the Lucene query.

3.3 Solver Engine

The *Solver Module* is responsible for finding a list of answers for a query. Each answer has a weight and a snippet: sentence or expression justifying the answer and it’s document identifier, as we can see in figure 6 for the question:

O que é um barrete frígio ? (What is a barrete frígio ?)

uma espécie de touca ou carapuça	93	doc:w07586 respSupport [96] O "barrete frígio" ou "barrete da liberdade" é uma espécie de touca ou carapuça, originariamente utilizada pelos moradores da Frigia (antiga região da Ásia Menor, onde hoje está situada a Turquia) (InfoEtno)
barrete da liberdade	92	doc:w07586 respSupport [96] O "barrete frígio" ou "barrete da liberdade" é uma espécie de touca ou carapuça, originariamente utilizada pelos moradores da Frigia (antiga região da Ásia Menor, onde hoje está situada a Turquia) (InfoEtno)
utilizada pelos sincretistas helenistas e romanos	91	doc:w07586 respSupport [96] O barrete frígio é utilizada pelos sincretismos(sincretistas helenistas e romanos, ainda que originalmente persa, deus salvador Mitra (divindade) (InfoEtno)

Fig. 6. Definition question result

The search for plausible answers is done on the Lucene selected documents by two tools: the *logic solver* and the *ad-hoc solver*. The semantic analyzer used before for the query will now create a DRS list for the selected texts. This list is a question dedicated Knowledge Base: the facts list. The *logic solver* is a logic-programming based module that performs a pragmatic interpretation of the query DRS over the full system knowledge base (the ontology and the facts list). It tries to find the best explanations for the question logic form to be true. This strategy for interpretation is known as “interpretation as abduction” [6].

The inference process is done with the Prolog resolution algorithm, which tries to unify the referents from the query with referents from documents, in the facts list, with help from the semantic information given by the ontology.

The *ad-hoc solver* is used for specific cases where a possible solution can be directly detected in the text. The system verifies each case conditions for the query and text expressions. Verifying the conditions might include a term semantic test for equivalence or “IsA” relation with another term, which is done by ontology analysis. Other conditions are related to text patterns, like ‘*X is DEFINITION*’, where the system attempts to learn the properties of X. This approach was used before in CLEF QA [8]. Figure 7 has a list of answers for the following question:

cerca de 960 km	91	doc w5109 respSupport [96] "Ceres" é um planeta anão que se encontra na cintura de asteroides , entre Marte e Júpiter . Ceres tem um diâmetro de cerca de 960 km e é o corpo mais maciço dessa região do sistema solar . contendo cerca de um terço do total da massa da cintura.(koi)
8 900 metros	91	doc w10034 respSupport [96] "Ceres" é um planeta anão com 8 900 metros de diâmetro.(koi)

Fig. 7. Numerical factoid question result

Qual o diâmetro de Ceres ? (What is the diameter of Ceres ?)

This is a *Factoid* question about a measure. The ad-hoc solver identified the term *diâmetro* (diameter) and searched for numerical answers, including the unit of measure (*km*, *metros*).

There are cases where several documents lead the system to a common answer. This is the case in figure 8, where the *ad-hoc solver* found two documents with the same temporal expression as an answer candidate to a *When* question. This enforces that answer’s weight.

The logic and ad-hoc found results are then merged to a final and weight sorted list. If the system finds more than one result for a question the QA@CLEF answer is the one with the maximum weight.

4 Results

In this QA@CLEF’s edition, the Universidade de Évora’s group registered for the monolingual Portuguese task, as did in previous participation [2], in 2005. A correct answer was found for 84 questions, which corresponds to an accuracy score of 42%.

Analyzing the results by question category, we can say that most of the errors were in the 90 wrong NIL returned values, where the system could not find an answer. Then, the *List* and *Temporally Restricted* questions represented a challenge and the obtained accuracy for these cases was around 20%. In the *Factoids* category the system had an accuracy close to the overall value, it was 39.62%. The best relative accuracy result was achieved in the *Definition* question type: 61.29%. Part of these definition answers were taken from Wikipedia documents, which sometimes had clear assertions. The overall Confidence Weighted Score over all assessed questions is 39.048/200 or 0.19524. All accuracy values for Portuguese as target are present in table 13 of QA Track Overview[9].

Comparing the current overall accuracy with the obtained in our department previous participation (25%) we believe this system produced good results. However, it needs some improvements as explained in the next section.

13 de maio de 1888	93	doc w8051 respSupport [95] A Lei Áurea foi assinada em 13 de maio de 1888, extinguindo a escravidão no Brasil.(h11_1_1,h11_1_1) doc w77212 respSupport [95] A escravidão só foi oficialmente abolida no Brasil com a assinatura da Lei Áurea . em 13 de maio de 1888 (h11_1_1)

Fig. 8. Temporal Expressions

5 Conclusions and Future Work

In this paper we describe our Question Answering System for QA@CLEF-2007. Compared with the system we used in 2005, the Senso system has a different methodology and is based on a different ontology.

Analyzing the incorrect answers, we saw that some questions had no candidate documents where to search for an answer. This means that the Lucene query used for document retrieval failed in those cases. In other cases of wrong NIL answers, the system could not find a possible answer in the retrieved documents.

Our semantic analyzer also had some problems with DRS generation, while analyzing the morpho-syntactical representation of non-trivial sentences. Other problems were related to incorrect pragmatic analysis, in the logic solver, due to ontology limitations and some lack of precision on the semantic information taken from the text sentences.

The Lucene search engine indexes all text collections and gives a list of documents that may have an answer and need detailed analysis. This was important to avoid a problem we had in 2005, related to time constraints, because some of the hard work is now done only over the selected documents. We need to correct the way the Lucene text search query is built, to fetch the answer candidate documents where it currently cannot do it.

We also intend to improve the Senso ontology. Since many operations in our methodology depend on it's content, it should be manually revised and extended. Along with this, some disambiguation tool would help for better precision when a sentence concept is being related with an ontology existent term.

The *ad-hoc solver* is a rule based answer generator. This participation in CLEF shows that more rules are needed and some of the existing ones need an adjustment.

In a future participation, we intend to apply our system to other source languages, with Portuguese as target language. This might require an extra question translation module.

References

1. Quaresma, P., Quintano, L., Rodrigues, I., Saias, J., Salgueiro, P.: The University of Évora approach to QA@CLEF-2004. In: CLEF 2004 Working Notes (2004)
2. Quaresma, P., Rodrigues, I.: A Logic Programming Based Approach To QA@CLEF05 Track. In: CLEF 2005 Working Notes (2005)
3. Saias, J., Quaresma, P.: A proposal for an ontology supported news reader and question-answer system. In: Rezende., S.O., et al. (eds.) 2nd Workshop on Ontologies and their Applications (WONTO 2006) in the Proceedings of International Joint Conference, 10th IBERAMIA, ICMC-USP, Ribeirão Preto, Brazil (2006) ISBN: 85-87837-11-7
4. Saias, J., Quaresma, P.: A methodology to create ontology-based information retrieval systems. In: Pires, F.M., Abreu, S. (eds.) EPIA 2003. LNCS (LNAI), vol. 2902. Springer, Heidelberg (2003)
5. Bick, E.: The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)

6. Hobbs, J., Stickel, M., Appelt, D., Martin, P.: Interpretation as abduction. Technical Report SRI Technical Note 499, 333 Ravenswood Ave., Menlo Park, CA 94025 (1990)
7. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer, Dordrecht (1993)
8. Tanev, H.: Extraction of Definitions for Bulgarian. In: CLEF 2006 Working Notes (2006)
9. Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: CLEF 2007 Working Notes (2007)

Web-Based Anaphora Resolution for the QUASAR Question Answering System

Davide Buscaldi, Yassine Benajiba, Paolo Rosso, and Emilio Sanchis

Dpto. de Sistemas Informáticos y Computación (DSIC)
Universidad Politécnica de Valencia, Spain
{dbuscaldi,ybenajiba,proso,esanchis}@dsic.upv.es

Abstract. This paper describes the work done by the RFIA group at the Departamento de Sistemas Informáticos y Computación of the Universidad Politécnica de Valencia for the 2007 edition of the CLEF Question Answering task. We participated in the Spanish monolingual task only. A series of technical difficulties prevented us from completing all the tasks we subscribed. Our 2006 system was modified in order to comply with the 2007 guidelines, especially with regard to anaphora resolution, tackled with a web based anaphora resolution module.

1 Introduction

QUASAR (QQuestion AnSwering And Retrieval) is the name we gave to our Question Answering (QA) system. It is based on the JIRS Passage Retrieval (PR) system [1], specifically oriented to this task. JIRS does not use any knowledge about the lexicon and the syntax of the target language, therefore it can be considered as a language-independent PR system. The system we used this year differs slightly from the one used in 2006. Its major improvement has been the insertion of an Anaphora Resolution module in order to comply with the guidelines of CLEF QA 2007. As evidenced in [2], the correct resolution of anaphora is crucial and allows to improve accuracy of more than 10% with respect to a system that does not implement any method for anaphora resolution. The web is an important resource for QA [3] and has been already used to solve anaphora in texts [4,5]. We took into account these works in order to build a web-based Anaphora Resolution module.

In the next section, we briefly describe the structure of our QA system, with particular emphasis on the new Anaphora Resolution module. In section 3 we discuss the results of QUASAR in the 2007 CLEF QA task.

2 System Architecture

In Fig.1 we show the architecture of the system used by our group at the CLEF QA 2007.

The user question is first examined by the *Anaphora Resolver* (AR). This module will pass the question to QUASAR in order to obtain the answer that

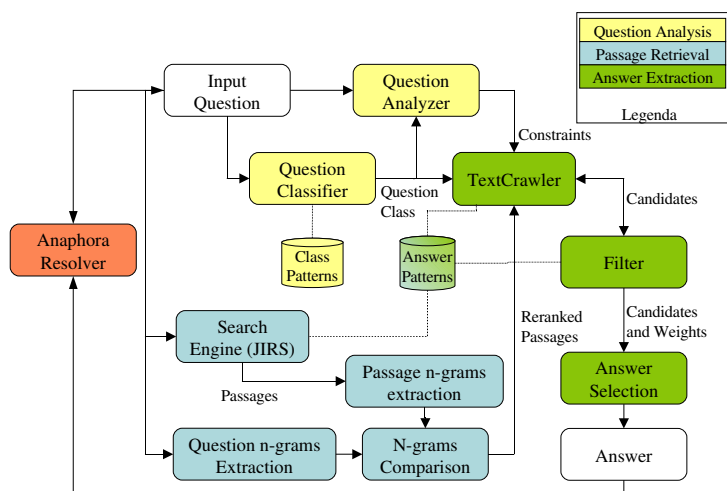


Fig. 1. Diagram of the QA system

will be used in the following questions of the same group (i.e., questions over the same topic) for the anaphora resolution. The AR hands over the eventually reformulated question 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.

A detailed description of each module goes beyond the scope of this paper. We describe in detail only the new AR module we added to the QUASAR system [6] in order to comply with the CLEF 2007 guidelines.

2.1 Anaphora Resolution Module

The anaphora resolution module carries out its work in 5 basic steps:

- Step 1: Using patterns in this step the module induces the different entity names, temporal and number expressions occurring in the first question of a group.
- Step 2 : For each question of the group the module replaces the entity names which occurred in the first question (which is most likely to contain the target) and occur only partially in the others. For instance, if the first question was “Cuál era el aforo del Estadio Santiago Bernabéu en los años 80?” (“What was the capacity of the Santiago Bernabeu stadium in the 80s?”) and another question within the same group was “Quién es el dueño del

estadio?” (“Who is the owner of the stadium?”), “estadio” in the second question is replaced by “Estadio Santiago Bernabéu”.

- Step 3: In this step the module focuses on pronominal anaphora and uses a web count to decide on whether to replace the anaphora with the answer of the previous question or with one of the entity names occurring in the first question of the group. For instance, if the first question was “Cómo se llama la mujer de Bill Gates?” (“What’s the name of the wife of Bill Gates?”) and the following question was “En qué universidad estudiaba él cuando creó Microsoft?” (“In which university was he studying when he created Microsoft?”), the module would check the Web counts of both “Bill Gates creó Microsoft” and “Melinda Gates creó Microsoft” and then it would replace the anaphora which would change the second question to “En qué universidad estudiaba Bill Gates cuando creó Microsoft?”.
- Step 4: The other type of anaphora left to be solved are the possessive anaphora. Similarly to the previous step, the module decides on how to change the question using web counts. For instance, if we take the same example mentioned in Step 2 and we say that a third question was “Cuánto dinero se gastó durante su ampliación entre 2001 y 2006?” (“How much money was spent for its enlargement between 2001 and 2006?”), the module would check the web counts of both “ampliación del Estadio Santiago Bernabéu” and “ampliación del Real Madrid Club de Fútbol” and change the question in order to become “Cuánto dinero se gastó durante ampliación del Estadio Santiago Bernabéu entre 2001 y 2006?”.
- Step 5: For the questions which have not been changed during any of the previous steps the module adds at the end of the question the entity name which has been found in the first question.

3 Experiments and Results

This year we experienced some technical difficulties with the result that we were able to submit only one run for the Spanish monolingual task. Moreover, we realized after the submission that JIRS indexed only a part of the collection, specifically the Wikipedia snapshot. The results is that we obtained only an accuracy of 11.5%. The reason of the poor performance is due partly to the fact that the patterns were defined for a collections of news and not for an encyclopedia, and partly to the fact that many questions had their answer in the news collection. For instance, in newspapers definitions are provided as expression between commas alongside with the entity being defined, whereas in an encyclopedia the entity is the title of the article and the definition is given in the first paragraph of the article. Moreover, Wikipedia includes templates for some categories of entities that contain most information. Currently our system is not able to process such templates. We obtained 54 NIL answers, more than the 25% of the total number of questions, and only once correctly. The Anaphora Resolution module did not perform particularly well, since we obtained only a 3.33% accuracy over linked questions, compared to an accuracy of 12.94% over self-contained questions.

4 Conclusions and Further Work

Our experience in the CLEF QA 2007 exercise was disappointing in terms of results, however we managed to develop an anaphora resolution module based on the web and we acquired valuable knowledge for our next participation. The first lesson we learnt from this participation was that the answers do not appear in the same form in Wikipedia and the newspaper collections. In order to address this problem we will need to implement different answer extraction methods, especially focused on encyclopedias, such as the one proposed in [7]. The second lesson was that the anaphora resolution method needs a deeper analysis of the question structure. We believe that syntactical parsing may improve the performance of this module.

Acknowledgments

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work.

References

1. Gómez, 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)
2. Vicedo, J.L., Ferrández, A.: Importance of pronominal anaphora resolution in question answering systems. In: ACL 2000: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, pp. 555–562. Association for Computational Linguistics (2000)
3. Lin, J.: The web as a resource for question answering: Perspectives and challenges. In: Language Resource and Evaluation Conference (LREC 2002), Las Palmas, Spain (2002)
4. Markert, K., Modjeska, N., Nissim, M.: Using the web for nominal anaphora resolution. In: EACL Workshop on the Computational Treatment of Anaphora, Budapest, Hungary (2003)
5. Bunescu, R.: Associative anaphora resolution: A web-based approach. In: EACL Workshop on the Computational Treatment of Anaphora, Budapest, Hungary (2003)
6. Buscaldi, D., Gómez, J.M., Rosso, P., Sanchis, E.: N-gram vs. keyword-based passage retrieval for question answering. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 377–384. Springer, Heidelberg (2007)
7. Fissaha, S., Jijkoun, V., de Rijke, M.: Fact discovery in wikipedia. In: 2007 IEEE/WIC/ACM International Conference on Web Intelligence. ACM Press, New York (2007)

A Lexical Approach for Spanish Question Answering

Alberto Téllez, Antonio Juárez, Gustavo Hernández, Claudia Denicia,
Esaú Villatoro, Manuel Montes, and Luis Villaseñor

Instituto Nacional de Astrofísica, Óptica y Electrónica

Laboratorio de Tecnologías del Lenguaje

Luis Enrique Erro no. 1, Sta. María Tonantzintla, Pue., Mexico

{albertotellezv, antjug, ghernandez, cdenicia,

villatoroe, mmontesg, villasen}@inaoep.mx

Abstract. This paper discusses our system's results at the Spanish Question Answering task of CLEF 2007. Our system is centered in a full data-driven approach that combines information retrieval and machine learning techniques. It mainly relies on the use of lexical information and avoids any complex language processing procedure. Evaluation results indicate that this approach is very effective for answering definition questions from Wikipedia. In contrast, they also reveal that it is very difficult to respond factoid questions from this resource solely based on the use of lexical overlaps and redundancy.

1 Introduction

Question Answering (QA) has become a promising research field whose aim is to provide more natural access to information than traditional document retrieval techniques. In essence, a QA system is a kind of search engine that allows users to pose questions using natural language instead of an artificial query language, and that returns exact answers to the questions instead of a list of documents.

Current developments in QA tend to use a variety of linguistic resources to help in understanding the questions and the documents. The most common linguistic resources include: part-of-speech taggers, parsers, named entity extractors, dictionaries, and WordNet [1,2,3]. In contrast to these developments that point to knowledge rich methods, we have proposed a straightforward QA system that avoids using any kind of linguistic resource, and therefore, that can be easily adapted to different domains and languages. This system is supported by two simple ideas. First, that questions and answers are commonly expressed using the same set of words, and second, that different kind of questions requires different kind of methods for their treatment.

A complete description of the proposed system can be found in [4]. This paper, on the contrary, focuses on discussing the system's evaluation results at the QA task of CLEF 2007. In particular, it gives some insights on the usefulness of lexical information for QA and also on the appropriateness of our approach for dealing with semi-structured collections such as Wikipedia.

2 Our System at a Glance

Our QA system is based on a full data-driven approach that exclusively uses lexical information to determine relevant passages as well as candidate answers. The system is divided in two basic components.

The first component focuses on answering *definition questions*. It determines the target term by a regular expression analysis, then it retrieves the most relevant page from Wikipedia using a traditional information retrieval technique, and finally, it extracts the target definition from the first paragraph of the selected page.

The second component focuses on answering *factoid questions*. It applies a passage retrieval process in order to find relevant passages from the EFE collection and Wikipedia. After that, it determines a set of candidate answers by a regular expression analysis. Finally, it uses a machine-learning strategy (a Naïve Bayes classifier) to calculate the confidence value for each candidate answer. In this case, the answer having the highest value is selected as final answer.

On the other hand, our system also contemplates the treatment of *linked questions*, where the first question indicates the focus of the group and the rest are somehow dependent from it. This treatment is quite simple: it basically considers the enrichment of dependent questions by adding some keywords (and the answer) from the self-contained question.

It is important to mention that this system continues our previous year work [5], but incorporates some new elements. Mainly, it takes advantage of the structure of Wikipedia to easily locate definition phrases, and applies a technique for query expansion based on association rule mining to enhance the passage retrieval (refer to [4] for more details).

3 Evaluation Results

This section presents the experimental results corresponding to our participation in the monolingual Spanish QA track at CLEF 2007. This evaluation exercise considered two basic types of questions, definition and factoid. However, this year also were included some groups of linked questions (where the first one –the self-contained question– indicates the focus of the group and the rest of them –the linked questions– are somehow dependent on it).

From the given set of 200 test question, our QA system treated 34 as definition questions and 166 as factoid. Table 1 details our general accuracy results. It is very interesting to notice that our method for answering definition questions was very precise. It could answer almost 90% of the questions; moreover, it never supplied wrong or unsupported answers. In addition, given that all these questions were answered from Wikipedia, this result evidenced that our approach could effectively take advantage of its inherent structure.

On the other hand, Table 1 also shows that our method for answering factoid questions was not completely adequate (it only could answer 23% of this kind of questions). Taking into account that 82% of the factoid questions were answered from Wikipedia, we presumed that the poor performance was caused

Table 1. System’s general evaluation

Questions	Right	Wrong	Inexact	Unsupported	Accuracy
Definition	30	–	4	–	0.88
Factoid	39	118	3	6	0.23
TOTAL	69	118	7	6	0.34

Table 2. Evaluation details about answering groups of linked questions

Questions	Right	Wrong	Inexact	Unsupported	Accuracy	NIL	
						Right	Wrong
Self-contained	64	95	6	5	0.38	3	35
Linked	5	23	1	1	0.17	0	5

by the Wikipedia’s structure. Two characteristics of Wikipedia damaged our system’s behavior. First, it is much less redundant than general news collections; and second, its style and structure favor the presence of anaphoric and ellipsis phenomena, and thus make lexical contexts of candidate answers less significant than those extracted from other free-text collections.

In order to illustrate the last problem consider the question *“How old was Alfred Hitchcock when he died?”*. A correct answer for this question is located at the Wikipedia’s document called *“Alfred Hitchcock”*, in the text fragment *“One year later, the April 29 of 1980, he died in his home located at The Angels when he was 80 years old ...”*. As can be noticed, the ellipsis in the text fragment (i.e., the omission of the name Alfred Hitchcock) produces a poor lexical overlap between the question and the answer’s context, and therefore, complicates the extraction of the given answer.

Finally, Table 2 shows some results from the treatment of groups of linked questions. It is clear that our approach was not useful for dealing with this kind of questions. The reason for this poor performance was that only 38% of the self-contained questions were correctly answered, and therefore, in the majority of the cases, the linked questions were enriched with erroneous information.

4 Conclusions

This paper presented a QA system that allows answering factoid and definition questions. This system is based on a lexical approach. Its main idea is that questions and their answers are commonly expressed using almost the same set of words, and therefore, it simply uses lexical information to identify relevant passages as well as candidate answers.

The proposed method for answering definition questions is quite simple; nevertheless it allowed achieving very high precision rates. We consider that its success is mainly attributable to its capability to take advantage from the style

and structure of Wikipedia (the used target document collection). On the contrary, our method for answering factoid questions was not equally successful. Paradoxically, the style and structure of Wikipedia, which favor the presence of anaphoric and ellipsis phenomena, caused a detriment in the lexical overlaps and in the answer redundancies, and consequently in the answer extraction process.

About the treatment of groups of linked questions, our conclusion is that the achieved poor performance (17%) was consequence of a cascade error. Only 38% of self-contained questions were correctly answered, and thus, most linked questions were expanded using incorrect information.

Acknowledgments

This work was done under partial support of CONACYT (project grant 43990 and scholarship 171610). We also thank the CLEF organizers.

References

1. de Pablo-Sánchez, C., González-Ledesma, A., Martínez-Fernández, J.L., Guirao, J.M., Martínez, P., Moreno-Sandoval, A.: MIRACLE's cross-lingual question answering experiments with Spanish as a target language. In: [6], pp. 488–491
2. Ferrés, D., Kanaan, S., Ageno, A., González, E., Rodríguez, H., Turmo, J.: The TALP-QA system for Spanish at CLEF. In: [6], pp. 400–409 (2005)
3. Roger, S., Ferrández, S., Ferrández, A., Peral, J., Llopis, F., Aguilar, A., Tomás, D.: AliQAn, Spanish QA system at CLEF 2005. In: [6], pp. 457–466
4. Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L.: INAOE's participation at QA@CLEF 2007. In: Working notes for the 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary (September 2007)
5. Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., Montes-y-Gómez, M., Villaseñor-Pineda, L.: Using machine learning and text mining in question answering. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 415–423. Springer, Heidelberg (2007)
6. 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. Springer, Heidelberg (2006)

Finding Answers Using Resources in the Internet

Mirna Adriani and Septian Adiwibowo

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
mirna@cs.ui.ac.id, septian.adiwibowo@ui.edu

Abstract. In this paper we describe our experiments in finding answers from documents based on statistical and linguistic knowledge. We collected the candidate answers from sources available on the internet, and then we used them to validate the answers found in the documents. The candidate answers from the documents were found using a statistical technique and linguistic knowledge such as named entity tags to find the type of answer that matches the question category.

1 Introduction

In our participation in the Question Answering task [1, 2] of Cross Language Evaluation Forum (CLEF) 2007, i.e., for Indonesian-English, we needed to use language resources to translate Indonesian queries into English. Luckily we found a machine translation tool available on the Internet that could be used to translate Indonesian queries into English.

We also made use of the information sources available on the Internet [3, 4] to validate answers that were found in the documents of a collection. We used statistical techniques to find the answers in the documents.

2 The Process of Analyzing the Questions

A number of steps were performed to the questions that we received from CLEF.

The query-answering process proceeds in the following stages:

1. Question categorization
2. Passages identification/building
3. Passages scoring
4. Answers identification.

First we categorize the Indonesian question according to the type of question. We identify the question type based on the question word found in the query.

The Indonesian question is then translated into English using a machine translation tool. The resulting English query is then used to retrieve relevant documents from the collection through an information retrieval system. The contents of a number of documents at the top of the list are then split into passages. The passages are then

scored using an algorithm, and the passage with the highest score is chosen to be the answer to the question.

2.1 Categorizing the Questions

Each question category, which is identified by the question word in the question, points to the type of answer that is looked for in the documents. We have 28 categories of question words based on the Indonesian grammar [5]. The Indonesian question-words used in the categorization are:

<i>dimana, dimanakah, manakah</i> (where)	points to <location>
<i>apakah nama</i> (what),	points to <location>
<i>siapa, siapakah</i> (who)	points to <person>
<i>berapa</i> (how many)	points to <measure>
<i> kapan</i> (when)	points to <date>
<i>organisasi apakah</i> (what organization)	points to <organization>
<i>apakah nama</i> (which)	points to <location> etc.

By identifying the question type, we can predict the kind of answer that we need to look for in the document. The Indonesian question is tagged using a question tagger that we developed according to the question word that appears in the question. This approach is similar to those used by Clark et al. and Hull [2, 3].

2.2 Building Passages

Next, the Indonesian question is translated into English using Toggletext¹, a machine translation system. The resulting English query is then run through an information retrieval system as a query to retrieve a list of relevant documents. We use Lemur² information retrieval system to index and retrieve the documents and passages. Passages are built from top 20 documents. Each passage contains 100 words. The passages are then tagged using GATE³ for Person, Location, Organization, Jobtitle, and Date. We also develop an additional tagger for Animal, Music-instruments, name-of-food, type-of-aircraft etc. The name entity tagger is developed based on the factoid information that we have collected from the internet.

2.3 Scoring the Passages

Passages are scored based on their likeliness to answer the question. The scoring rules consider the number of words from the questions that appear on the passages. Then the average distance weight of the answer candidates and the words that appear on the query are also considered.

$$\text{Average Distance Weight} = \left(\sum_{i=0}^n \text{Distance}(W, q_i) \right) / n$$

¹ See <http://www.toggletext.com/>

² See <http://www.lemurproject.org/>

³ See <http://www.gate.shef.ac.uk/>

where

N	: the number of words in a question
q_i	: the i -th word that appears in a question
W	: a candidate answer which has the matching entity-name tag that appears in a passage
distance	: a number of words that appear between W and q_i .

Once the passages obtained their scores, the top 20 passages with the highest scores and have the appropriate tags – e.g., if the question type is person (the question word “*who*”) then the passages must contains the person tag – are then taken to the next stage.

2.4 Finding the Answer

The top 20 passages are analyzed to find the best answer. The likeliness of a word to be the answer to the question is inversely proportional to the number of words in the passage that separate the candidate word and the word in the query. For each word, its distance from a query word found in the passage is computed. The candidate word that has the smallest distance is the final answer to the question. We also validate the answer candidates to the answer that we find on available sources on the internet. We get the top 50 snippets for each question from Google (<http://www.google.com>). We then rank the words according to their word frequencies. The word that has the highest frequency is the answer candidate to a question. We then add a weight to the final score of the answer found in the document. The final score of the answer is the sum of the score derived from Google and Average Distance Weight. The value of the final score is achieved by assigning proportions to the Google-based score and the average distant weight score. The values of the proportions are set through preliminary experimentation.

$$F = a G + b ADW$$

where

F	: the final score of a candidate answer
G	: word frequency that appears in Goggle’s snippet
ADW	: the average distance weight of a candidate answer
a, b	: parameters representing proportions, where $a + b = 1.0$.

For the definition questions, we employ Apple Pie Parser (http://nlp.cs_nyu.edu/app) to do constituency parsing for the passages. The answer is extracted from noun-phrases (mostly in form of apposition) from the top 20 passages. The query words are deleted from the noun phrase.

3 Experiment

We participated in the bilingual task with English topics. The query translation process was performed fully automatic using a machine translation technique. The machine translation technique translates the Indonesian queries into English using Toggletext⁴, a machine translation that is available on the Internet. In these experiments, we used

⁴ See <http://www.toggletext.com/>

Lemur⁵ information retrieval system which is based on the language model to index and retrieve the documents.

4 Results

Our work is focused on the bilingual task using Indonesian questions to retrieve answer from an English document collection. Table 1 shows the result of our experiments.

Table 1. The QA results

Task : Bilingual QA	Evaluation
W (wrong)	175
U (unsupported)	1
X (inexact)	4
R (right)	20

Changes in the question types this year had an impact on the number of answers that we managed to find. We have developed our scoring and answer patterns using the previous year's questions. However, they did not work very well for this year's questions. The percentage of correct answers that we got this year was only 10%.

5 Summary

We learned from our work that using information from sources available on the internet can help verify the answers found in documents. However, deeper linguistic knowledge such as syntax needs to be considered to get an even better result. We had used a syntactic parser for Indonesian Question Answering System that gave positive result, so we plan to use the same approach for English Question Answering in the future.

References

1. Clarke, C.L.A., Cormack, G.G., Kisman, D.I.E., Lynam, K.: Question Answering by Passage Selection: The 9th Text retrieval Conference (TREC-9) (2000)
2. Hull, D.: Xerox TREC-8 Question Answering Track Report: The 8th Text Retrieval Conference (TREC-8) (1999)
3. Hildebrandt, W., Katz, B., Lin, J.: Answering definition questions with multiple knowledge sources. In: Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL 2004) (2004)
4. Hovy, E., Hermjakob, U., Lin, C.Y.: The use of external knowledge in factoid QA. In: Proceedings of the Tenth Text REtrieval Conference (TREC 2001), pp. 644–652 (2001)
5. Alwi, H., et al.: Indonesian Grammar. Center of Bahasa Indonesia, Jakarta (1998)

⁵ See <http://www.lemurproject.org/>

UAIC Romanian QA System for QA@CLEF

Adrian Iftene¹, Diana Trandabăț^{1,2}, Ionuț Pistol¹, Alex Moruz^{1,2}, Alexandra Balahur¹,
Diana Cotelea¹, Iustin Dornescu¹, Iuliana Drăghici¹, and Dan Cristea^{1,2}

¹ UAIC: Faculty of Computer Science, “Alexandru Ioan Cuza” University, Romania

² Institute for Computer Science, Romanian Academy Iasi Branch

{adiftene, dtrandabat, ipistol, amoruz, abalahur, dcotelea,
idornescu, idraghici, dcristea}@info.uaic.ro

Abstract. This paper briefly describes UAIC¹’s participation in this year’s CLEF question answering competition, focusing on the main challenges and changes compared to our last year participation. An analysis of the errors introduced by each module is also discussed.

Keywords: Question Answering, Anaphora Resolution, Error Analysis.

1 Introduction

The first version of our QA system was developed as part of our participation in the last year’s competition (Pușcașu et al., 2006) where we took part in the RO-EN track (the first language represents the source language of the questions and the second represents the language of the document collection in which the system searches for the answer). In the 2007 competition, organizers introduced several tasks that included Romanian: RO-RO, EN-RO, RO-EN. The Romanian corpus used in this competition was the Romanian Wikipedia, frozen at the level of November 2006.

The general system architecture and the most important modules are described in the next section, with special focus on newly inserted components. The third section analyse the errors and in the last section we describe some plans to further develop our system.

2 QA System Architecture

2.1 Corpus Pre-processing

The set of available Wikipedia documents includes 180.500 files with a total size of 1.9 GB. The documents include topic related information, as well as forum discussions, images and user profiles.

In order to transform the 1.9 GB corpus into a format more manageable by the available processing tools, two pre-processing steps were performed:

- Documents considered irrelevant for the answer extraction task were removed from the set. These documents include images, user profiles and forum

¹ “Alexandru Ioan Cuza” University.

discussions. The filtering was performed automatically using a pattern based cleaning method.

- In a second step, all remaining documents were converted to a simpler format by stripping off most of the *html* markups. The only markups kept are the name of the article and those indicating paragraphs.

After these two filtering steps the size of the corpora was reduced down to approximately 175 MB. This significant reduction contributed to making the linguistic processing steps more time-efficient.

Both the Wikipedia corpus and the set of test questions went through the same linguistic processing steps prior to the actual analysis: tokenization, POS-tagging, lemmatization and NER for the types Person, Location, Measure, Organization and Date.

The linguistic processing steps required significant processing power, as both the size of the corpus and the number of tools involved were quite large. On a single computer the processing time for the entire corpus was estimated at 6 to 8 hours. In order to reduce this interval, to be able to run more improvement loops, we have used a multi-processor computer capable of running 10 parallel processes. With a minimal adaptation of the tools, we were able to reduce the processing time to about 30 minutes per loop.

2.2 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 focus, the question type and the set of keywords relevant for the question. To achieve these goals, our question analyzer performs the following steps:

i) NP-chunking, Named Entity extraction

Using the pre-processed set of questions as input, on the basis of the morpho-syntactic annotation provided by our tools, a rule-based shallow noun phrase identifier was implemented. The same NE recognizer employed during the pre-processing of the corpus provides the set of NEs in the question.

ii) Question focus

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", as in "*What is Wikipedia?*".

iii) The answer type

Our system is able to identify the following types of answers: Person, Location, Organization, Temporal, Measure, Definition and Other. The assignment of a type to an analyzed question is performed using specific patterns for every type. In the case of ambiguous questions (e.g. *What*), the answer type is computed starting from the question focus. We extracted from WordNet a set of lists with all hyponyms of the first five² answer types and tried to identify the question focus among elements of these

² Definition type is computed only by pattern matching methods and Other type is assigned to questions that didn't enter in any other category.

lists. For example, in the case of the question *What city is usually identified with the Homeric Troy?*, the question focus is city, noun found in the Location list, therefore the associated expected answer type is Location. Thus, the answer of this question is searched in a database set of city names.

iv) Inferring the question type

Question type can be: Factoid, List or Definition. In order to identify the question type we used three simple rules: if the expected answer type is Definition, then the question type is definition; if the question focus is a plural noun, then the question type is list, otherwise it is factoid.

v) Keyword generation

The set of keywords is composed of the question focus, the identified NEs in the question, nouns, adjectives, and non-auxiliary verbs belonging to the question. For example, for the first question in Table 1, the keywords list is {*faimos, romancier, novelist, realizator, american, a trăi, 1899, 1961*} (En: {*famous, novelist, short-story writer, producer, American, to live, 1899, 1961*}).

Table 1. Example of questions grouped in a topic

<Group id=1>	
<Question id=1>	<i>Ce faimos romancier, nuvelist și realizator american de povestiri a trăit între anii 1899 și 1961?</i> ³
<Question id=2>	<i>Pentru ce premiu a fost el laureat în anul 1954?</i> ⁴
<Question id=3>	<i>În ce an a fost el laureat al Premiului Pulitzer?</i> ⁵
</Group>	

vi) Anaphora resolution

Every year new features are added in the QA@CLEF competitions. This year the new feature was the grouping of questions into series. All questions belonging to a series address the same general topic, the domain of which is usually specified by either the first question of the series or by its answer. Mainly, questions of a series are tightly coupled by anaphoric links that involve entities mentioned in previous questions or their answers.

The new feature for this year’s QA@CLEF competition, grouping questions on topics (Table 1), requires the addition of a new module in the general architecture of the system, responsible for anaphora resolution. We can see that in questions 2 and 3 the pronoun “*el*” (En: “*he*”) must be replaced with the answer of the first question. For solving this problem we adopted two methods of anaphora resolution presented below:

1. Antecedent identification

We classified the possible anaphoric relations into the following cases:

- questions with anaphors that refer to a previous question answer;
- questions with anaphors that refer to a previous question focus.

³ En: *What famous American novelist and short story writer lived between 1899 and 1961?*

⁴ En: *Which prize was he nominated for in 1954?*

⁵ En: *What year was he nominated for the Pulitzer Prize?*

Empirical methods were used to decide if the focus or the answer of the first question in a group is needed.

For the example in table 1, the anaphora resolution module decided that the answer of the first question in group (“*Ernest Hemingway*”) should be added to the keywords of the second question, the set of keywords becomes: {*premiu, laureat, 1954, Ernest Hemingway*}.

2. The Backup solution

Since the first method depends on the capability of our system to identify correctly the answer, we considered also a second method. The second solution was adding all keywords of the first question in the group to the keywords of the other questions. For example, the keywords list for the second sentence in the group presented in table 1 becomes {*faimos, romancier, nuvelist, realizator, american, a trăi, 1899, 1961, premiu, laureat, 1954*}.

2.3 Index Creation and Information Retrieval

The purpose of this module is to retrieve the relevant snippets of text for every question. Below is a brief description of the module:

i) Query creation

Queries are created using the sequences of keywords and Lucene⁶ mandatory operator “+”. In this manner we obtain a regular expression for every question, which is then used in the search phase. The queries have also been enhanced by the addition of the question focus and synonyms for all the keywords. For example, the query attached to question 1 in the test data (from table 1) will be:

+romancier (faimos renumit vestit) (nuvelist nuvelistic) (realizator înfăptuitor) american (trăi viețuit) 1899 1961

If a word is not preceded by any operator, then that word is optional. The words between the brackets are connected by the logical operator OR, which means that at least one of these elements should be found in a snippet in order for it to be returned by the retrieval module.

ii) Index creation

We have created the index of the document collection using the lemmas determined in the pre-processing phase. We have created two indexes, one at paragraph level and one at document level.

Index creation at the paragraph level

The main purpose of this type of indexing is to identify and retrieve a minimum useful amount of information related to a question. This method’s drawback was that, in some cases, a paragraph was just a single phrase. Of course the advantage is that from a reduced amount of information, we could easier identify and extract the answer from the retrieved paragraph.

⁶ <http://lucene.apache.org/>

Index creation at document level

An alternative indexing method was indexing at article level. The disadvantage of this method came when we had to extract the answer, when more refined algorithms were necessarily.

iii) Relevant paragraph extraction

Using the queries and the index, we extracted with Lucene a ranked list of articles / paragraphs for every question.

2.4 Answer Extraction

The retrieving process depends on the expected answer type: the answer retrieval module identifies the named entities in every snippet provided by Lucene and matches them to the answer type. When the answer type is not an entity type name, the answer retrieval syntactic patterns are based on the question focus.

3 System Performance and Errors Analysis

The official evaluation of the CLEF@QA 2007 competition revealed the results presented in table 2 for the Romanian system:

Table 2. Official results

Result evaluation		
Z	UNKNOWN	0
R	CORRECT	24
U	UNJUSTIFIED	1
W	WRONG	171
X	INEXACT	4
TOTAL		200

Each answer was evaluated as being UNKNOWN (unevaluated), CORRECT, UNJUSTIFIED (no supporting snippet provided), WRONG or INEXACT (incomplete answer). The precision of our system was 12%, similar to the accuracy obtained last year. However, seen the greater complexity of the task as compared to the previous year, we may estimate that the performance of the system improved.

We considered two methods in order to perform the system error analysis:

- Starting from the incorrect answered questions, going backwards and trying to identify what went wrong.
- Performing error analysis per module.

The analysis presented in this section is mainly derived from the first method.

Corpus pre-processing errors

The major problem of the pre-processing phase was related to the incomplete cleaning of the HTML format of Wikipedia. The most common problems left unsolved in

Table 3. Number of errors introduced per module

Module	Submodule	% of Errors	Errors Number
<i>Corpus Pre-processing</i>	HTML Cleaning	10	20
	Answer type identification	14	28
<i>Question Analysis</i>	NER	5	11
	Anaphora resolution	11	23
	Keyword generation	7	14
<i>Indexing and retrieval</i>	Query building	11	23
	Retrieval	3	6
<i>Answer extraction</i>	Answer extraction	40	80

HTML cleaning were keeping titles of articles that are referred within a Wikipedia entry, as well as empty Wikipedia articles. Because, once indexed at paragraph level, confused the TF/IDF ranking measure of the search module (since they contain only few words, among which usually many included in the query). Another problem was the existence, in some indexed paragraphs, of formatting tags, which increased the challenge for the answer extraction module.

Question analysis errors

The main errors introduced by the question analysis are mainly due to:

- incorrect answer type identification,
- name entity misrecognition,
- improper generation of keywords list,
- incorrect anaphora resolution.

The first class of errors is due to incorrect answer type identification from questions. We can see in table 4 that most misidentified answer type occurred for the Organization type. The reason is that our NER module didn't identify organizations like "Eteria", "Ansamblul General al Națiunilor Unite" (En: United Nations General Assembly) etc.

Table 4. Example of errors introduced by answer type identification

Answer Type	Correct	Incorrect	Total	Incorrect Percent
Person	34	6	40	15 %
Time	16	1	17	6 %
Other	55	3	58	5 %
Organization	12	13	25	52 %
Count	35	1	36	3 %
Location	20	4	24	17 %
	172	28	200	14 %

The second class of errors is due to compound nouns and named entities. Thus, the groups "Traian Băsescu", "Agenția de Securitate Națională" (En: National Security Agency), "Empire State Building", "Al doilea război mondial" (En: Second World

War) are not recognized by the NER module and treated as normal composed sequences, thus eliminating stop words, inflecting adjectives, etc.

Mainly, the errors produced by the keyword generation module are due to data sparseness. For each noun phrase or verb considered for inclusion in the keywords list, synonyms and inflexions were generated. If the question contained many nominal groups, the list increases considerably. Thus, many logical operators have to be used, so that the retrieval module returns too much noise.

For instance, for question 197 “Care este mărimea a cărei unitate de măsură are același nume ca și un limbaj de programare?” (En: “Which is the measure whose unit has the same name as a programming language?”), the noun phrases in the question (measure, unit, programming language, name) are too general and too common in the document collection. So that a more refined pattern needs to be found for those cases. This kind of errors is usually linked to query building, since considering very general noun phrases as mandatory for the query generates lots of irrelevant documents.

The opposite situation was a misinterpretation of superlatives in the keywords list generation: “cel mai mare producător de telefoane mobile” (En: “biggest mobile telephone producer”), “cel mai înalt vârf din Kilimanjaro” (En: “the highest peak in Kilimanjaro”). The superlative “cel mai” was considered a stop word and deleted from the list of keywords. A similar situation was encountered in question 43, where the restriction “la început” (En: “at the beginning”) was lost.

Indexing and information retrieval

The main problem of the query building was the recognition of the focus, the word of the question that bears clues that help finding the answer of the question. This is usually following a *wh*- word or is a noun indicating the answer’s hypernymy class.

In only 102 questions of the 200, the extracted snippet contained the answer (“found” cases from Figure 1).

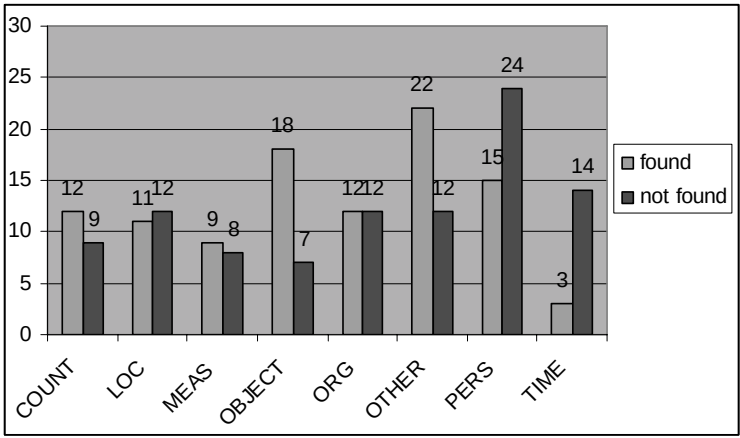


Fig. 1. Gold answer in extracted snippets

Answer extraction

The answer extraction module caused the biggest error rate, 40%. The main causes are that it relies too much on the NER, the patterns used for answer extraction were incomplete, and our patterns always select only the first entity corresponding to the desired type of the answer, this way paying no attention to other possibilities. Table 5 presents the situations (ordered by consecutive snippets) where the answer was present in the snippet, and was not recovered by the answer extraction module.

Table 5. Answer extraction accuracy

Answer was in snippet no.	1	2	3	4	5	6-8	9-20
Answer not extracted (%)	53.75	15	6.25	7.5	3.75	2.5	1.25

4 Conclusions

This paper presents the Romanian Question Answering system which was enrolled and participated in CLEF 2007. The official evaluation showed an overall accuracy of 12%, which, although lower than the other Romanian systems (Tufiş et al., 2007), indicates similarities with the other systems participating at QA@CLEF (Giampiccolo et al., 2007) and an important start-up for a Wikipedia based Romanian Question Answering system. An analysis of the errors of our system is intended to help us to enhance the system to be presented at the next year's competition.

As a further development, in order to check if the found candidate term is valid, and thus, if the answer formulated to the question is correct, we intend to use a Textual Entailment (TE) module. In order to use the textual entailment system, we built patterns to transform the questions with the answer type PERSON, LOCATION, DATE and ORGANIZATION in statements. From the tests performed, we noticed that using a TE module within the QA system results in the improvement of the ranking among possible answers. This is possible due to the fact that the TE system performs a deep semantic analysis of the question context and does not solely apply lexical distances among words.

References

- Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Săcăleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 19-21 September (2007)
- Puscasu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceaşu, A., Ştefănescu, D., Ion, R., Dornescu, I., Moruz, A., Cristea, D.: Cross-Lingual Romanian to English Question Answering at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 385–394. Springer, Heidelberg (2007)
- Tufiş, D., Ştefănescu, D., Ion, R., Ceaşu, A.: RACAI's Question Answering System at QA@CLEF 2007. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 19-21 September (2007)

The University of Amsterdam's Question Answering System at QA@CLEF 2007

Valentin Jijkoun, Katja Hofmann, David Ahn, Mahboob Alam Khalid,
Joris van Rantwijk, Maarten de Rijke, and Erik Tjong Kim Sang

ISLA, University of Amsterdam

{jijkoun,khofmann,ahn,mahboob,rantwijk,mdr,erikt}@science.uva.nl

Abstract. We describe a new version of our question answering system, which was applied to the questions of the 2007 CLEF Question Answering Dutch monolingual task. This year, we made three major modifications to the system: (1) we added the contents of Wikipedia to the document collection and the answer tables; (2) we completely rewrote the module interface code in Java; and (3) we included a new table stream which returned answer candidates based on information which was *learned* from question-answer pairs. Unfortunately, the changes did not lead to improved performance. Unsolved technical problems at the time of the deadline have led to missing justifications for a large number of answers in our submission. Our single run obtained an accuracy of only 8% with an additional 12% of unsupported answers (compared to 21% in the last year's task).

1 Introduction

For our earlier participations in the CLEF question answering track (2003–2006), we have developed a parallel question answering architecture in which candidate answers to a question are generated by different competing strategies, *QA streams* [4]. Although our streams use different approaches to answer extraction and generation, they share the mechanism for accessing the collection data: we have converted all of our data resources (text, linguistic annotations, and tables of automatically extracted facts) to fit in an XML database in order to standardize the access [4]. For the 2007 version of the system, we have focused on three tasks:

1. Add to the data resources of the system material derived from the Dutch Wikipedia (previously only derived from Dutch newspaper text).
2. Rewrite the out-of-date code which takes care of the communication between the different modules (previously in Perl) in Java. In the long run we are aiming at a system which is completely written in Java and is easily maintainable.
3. Add a new question answering stream to our parallel architecture: a stream that generates answers from pre-extracted relational information based on *learned* associations between questions and answers; a similar stream in last year's system used manual rules for identifying such associations.

This paper is divided in seven sections. In section 2, we give an overview of the current system architecture. In the next three sections, we describe the changes made to our system for this year: resource adaptation (section 3), code rewriting, and the new table stream (4). We describe our submitted runs in section 5 and conclude in section 6.

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. The answer extraction streams can be divided into three groups based on the text corpus that they employ: the CLEF-QA corpus, Dutch Wikipedia, or the Web. Below, we describe these briefly.

The Quartz system (Figure 1) contains four streams that generate answers from the two CLEF data sources, the CLEF newspaper corpus and Wikipedia. The *Table Lookup* stream searches for answers in specialized knowledge bases which are extracted from the corpus offline (prior to question time) by predefined rules. These information extraction 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. The stream uses the analysis of a question to determine how a candidate answer should be looked up in the database using a manually defined mapping from question to database queries. Our new stream, *ML Table Lookup*, performs the answer lookup task by using a mapping learned automatically from a set of training questions (see section 4 for a more elaborate

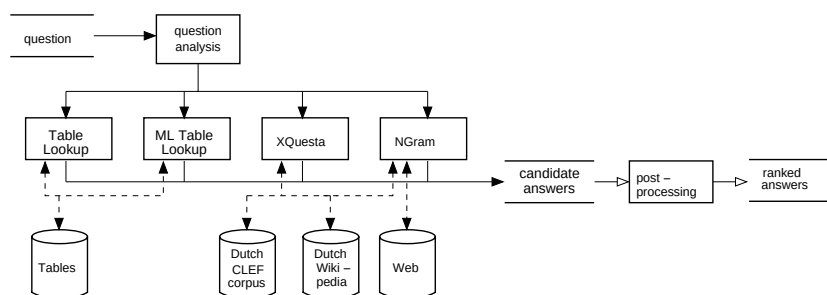


Fig. 1. Quartz-2007: the University of Amsterdam's Dutch Question Answering System. After question analysis, a question is forwarded to two table modules and two retrieval modules, all of which generate candidate answers. These four question processing streams use the two data sources for this task, the Dutch CLEF newspaper corpus and Dutch Wikipedia, as well as fact tables which were generated from these data sources, and the Web. Related candidate answers are combined and ranked by a postprocessing module, which produces the final list of answers to the question.

description). The *Ngram* stream looks for answers in the corpus by searching for most frequent word ngrams in a list of text passages retrieved from the collection using a standard retrieval engine (Lucene) using a text query generated from the question.

The most advanced of the four streams is XQuesta. For a given question, it automatically generates XPath queries for answer extraction, and executes them on an XML version of the corpus which contains both the corpus text and additional annotations. The annotations include information about part-of-speech, syntactic chunks, named entities, temporal expressions, and dependency parses (from the Alpino parser [6]). For each question, XQuesta only examines text passages relevant to the question (as identified by Lucene).

There is one stream which employs textual data outside the CLEF document collection defined for the task: the *Ngram* stream also retrieves answers by submitting automatically generated web queries to the Google web search engine and collecting most common ngrams from the returned snippets. The answers candidates found by this approach are not backed up by documents from the CLEF collection as required by the task. For this reason such candidates are never returned as actual answers, but only used at the answer merging stage to adjust the ranking of answers that are found by other QA streams.

3 Wikipedia as a QA Resource

Our system uses Dutch Wikipedia in the same way as the Dutch newspaper corpus. We used an XML dump of Wikipedia¹ that provides basic structural markup and additionally annotated it with sentence boundaries, part-of-speech tags, named entities and temporal expressions. Wikipedia was consulted by the XQuesta and NGram streams and was also used for offline information extraction.

4 Machine Learning for QA from Tabular Data

As described in section 2, our offline information extraction module creates a database of simple relational facts to be used during question answering. A *TableLookup* QA stream uses a set of manually defined rules to map an analyzed incoming question into a database query. A new stream, *MLTableLookup*, uses supervised machine learning to train a classifier that performs this mapping. In this section we give an overview of our approach. We refer to [5] for further details.

Essentially, the purpose of the table lookup stream is to map an incoming question to an SQL-like query “select AF from T where $sim(QF, Q)$ ”, where T is the table that contains the answer in field AF and its other field QF has a high similarity with the input question Q . Executing such a query for a given question results in a list of answer candidates—the output of the *MLTableLookup* stream.

¹ URL: <http://ilps.science.uva.nl/WikiXML>

In the query formalism described above, the task of generating the query can be seen as the task of mapping an incoming question Q to a triple $\langle T, QF, AF \rangle$ (a *table-lookup label*) and defining an appropriate similarity function $\text{sim}(QF, Q)$.

The database of facts extracted from the CLEF QA collection consists of 16 tables containing 1.4M rows in total. For example, the *Definitions(name, definition)* table contains the definition of *Soekarno* as *president of Indonesia*, the table *Birthdates(name, birthdate)* contains the information that *Aleksandr Poesjkin* was born in 1799. Then, for a question such as *Wie was de eerste Europese commissaris uit Finland?* (*Who was the first European Commissioner from Finland?*) the classifier may assign the table-lookup label $\langle T : \text{Definitions}, QF : \text{Definition}, AF : \text{name} \rangle$. In this case, the question would be mapped to the SQL like query "select name from *Definitions* where $\text{sim}(\text{definition}, \{\text{eerste}, \text{European}, \text{commissaris}, \text{Finland}\})$ ".

For an incoming question, we first extract features and apply a statistical classifier that assign a *table-lookup label*, i.e., a triple $\langle T, QF, AF \rangle$. We then use a retrieval engine to locate values of field QF in table T which are most similar to the text of the question Q (according to a retrieval function $\text{sim}(\cdot, \cdot)$), and return values of corresponding AF fields. Figure 2 shows the architecture of our system.

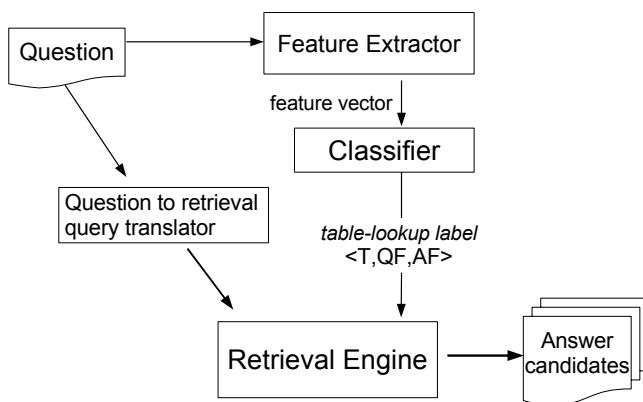


Fig. 2. Architecture of the MLTableLookup QA stream

Our architecture depends on two modules: the classifier that predicts table lookup labels and the retrieval model $\text{sim}(\cdot, \cdot)$ along with the text representation and the retrieval query formulation. For the later task we selected Lucene's vector space model as retrieval model, and used a combination of two types of text representation, exact and stemmed forms of the question words, to formulate a retrieval query, i.e., to translate an incoming question to a retrieval query.

The interesting and novel part of the new QA system is the second stage of our query formulation, i.e., training a classifier to predict table lookup labels. This stage, in turn, can be split in two parts: generating training data and actually training the classifier. We generate the training data using the selected retrieval

model. We index the values of all fields of all rows in our database as separate documents. For each question q , we translate the question into the retrieval query and use the selected retrieval model to generate a ranked list of field values from our database. We select the document, table name T and the field name QF , such that it occurs first time in the ranked list and the value of some other field AF contains the answer of the question. In other words we find T , QF and AF such that the query “select AF from T where $sim(QF, Q)$ ” returns a correct answer to question q at the top rank. We use the label $\langle T, QF, AF \rangle$ as a correct class for question q . For example, we translate the question *In welk land in Afrika is een nieuwe grondwet aangenomen?* (whose answer is *Zuid-Afrika*) into a retrieval query that is composed of the question words and words retrieved from the process of filtering out stopwords and stem the remaining the question words. Then we run the query against the retrieval engine’s index; for this particular example our system finds the triplet $\langle T : Locations, QF : location_b, AF : location_a \rangle$ as the optimal table-lookup label for this question.

Next, in order to generate training data, we represent each question as a set of features. We use the existing module of [2] to construct the set of features. Finally we train a memory-based classifier TiMBL [1] and use a parameter optimization tool to find the best setting for Timbl; see Figure 3 for an overview.

We used a set of question/answer pairs from the CLEF-QA tasks 2003–2006 and a knowledge base with tables extracted from the CLEF-QA corpus using the information extraction tools of QUARTZ system. We split our training corpus of 644 questions with answers into 10 sets and run a 10-fold cross-validation. The performance of the system is measured using the Mean Reciprocal Rank (MRR, the inverse of the rank of the first correct answer, averaged over all questions) and accuracy at n ($a@n$, the number of question answered at rank $\leq n$). Table 1 shows the evaluation results averaged over the 10 folds.

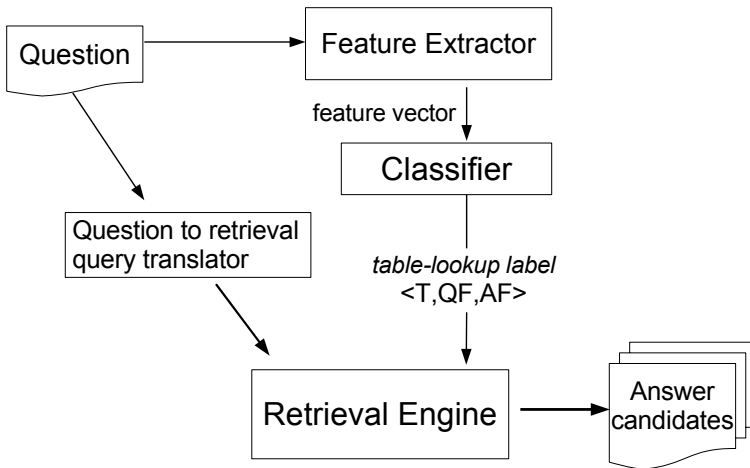


Fig. 3. Learning table lookup labels

Table 1. Evaluation of the ML Table-lookup QA stream applied to the CLEF 2003–2006 question answer pairs

a@1	a@5	a@10	MRR
13.1%	21.4%	24.1%	0.593

5 Runs

We have submitted a single Dutch monolingual run: uams071qrtz. The associated evaluation results can be found in Table 2. In this run, we have treated list questions as factoid questions: always returning the top answer. The planned updates of the system proved to be more time consuming than was expected. The system was barely finished at the time of the deadline. Because of this there was no time for an elaborate test or for compiling alternative runs. The performance of the system has suffered from this: only about 8% of the questions were answered correctly. The previous version achieved 21% correct on the CLEF-2006 questions.

The prime cause of the performance drop can be found in the submitted answer file. No less than 81 (41%) of the 200 answers did not contain the required answer snippet. This problem caused all but 4 of the unsupported assessments. 22 of these 81 answers mentioned a document id for the missing snippet but the other 59 lacked the id as well. The problem was caused by a mismatch between the new java code and the justification module which caused all justifications associated with answers from the two table streams to be lost.

When examining the answers for the factoid and definition questions, we noticed that a major problem is a mismatch between the expected answer type and the type of the answer. Here are a few examples:

0003. How often was John Lennon hit? Answer: Yoko Ono
 0136. What is an antipope? Answer: Anacletus II
 0160. Who is Gert Muller? Answer: 1947

Table 2. Assessment counts for the 200 top answers in the Amsterdam run submitted for Dutch monolingual Question Answering (NLNL) in CLEF-2007. About 8% of the questions were answered correctly. Another 12% were correct but insufficiently supported. The run did not contain NIL answers.

Question type	Total	Right	Unsupported	Inexact	Wrong	% Correct
factoid	156	14	17	0	125	9%
definition	28	1	5	0	22	4%
factoid+definition	184	15	22	0	147	8%
list	16	0	1	1	14	0%
temporarily restricted	41	2	3	0	36	5%
unrestricted	159	13	20	1	125	8%
all	200	15	23	1	161	8%

As many as 61 of the 161 incorrectly answered displayed such a type mismatch. The question classification part of the system (accuracy: 80%) generates an expected type for each answer but it is not used in the postprocessing phase. Indeed, the addition of a type-based filter at the end of the processing phase is one of the most urgent tasks for future work.

6 Conclusion

We have described the fifth iteration of our system for the CLEF Question Answering Dutch mono-lingual track (2007). While keeping the general multi-stream architecture, we re-designed and re-implemented the system in Java. This was an important update, which however did not lead to improved performance, mainly due to many technical problems that were not solved by the time of the deadline. In particular, these problems led to the loss of originating snippets for many of the answer candidates extracted from the collection, resulting in a large number of unsupported answers in our submission. Our single run obtained an accuracy of only 8% with an additional 12% of unsupported answers (last year, our best run achieved 21%).

Addressing these issues, performing a more systematic error analysis and answer extraction step in XQuesta stream and learning step in MLTableLookup are the most important items for future work.

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, and 640.002.501. Mahboob Alam Khalid and Katja Hofmann were supported by NWO under project number 612.066.512.

References

1. 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/>
2. Jijkoun, V., Mishne, G., de Rijke, M.: Building infrastructure for Dutch question answering. In: Proceedings DIR 2003 (2003)
3. 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)

4. Jijkoun, V., van Rantwijk, J., Ahn, D., Sang, E.T.K., de Rijke, M.: The University of Amsterdam at QA@CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
5. Khalid, M.A., Jijkoun, V., de Rijke, M.: Machine learning for question answering from tabular data. In: FlexDBIST 2007 Second International Workshop on Flexible Database and Information Systems Technology (2007)
6. van Noord, G.: At last parsing is now operational. In: Proceedings of TALN 2006, Leuven, Belgium (2006)

Combining Wikipedia and Newswire Texts for Question Answering in Spanish^{*}

César de Pablo-Sánchez¹, José L. Martínez-Fernández^{1,2},
Ana González-Ledesma³, Doaa Samy¹, Paloma Martínez¹,
Antonio Moreno-Sandoval³, and Harith Al-Jumaily¹

¹ Universidad Carlos III de Madrid

{cdepablo,dsamy,pmf,haljumai}@inf.uc3m.es

² DAEDALUS, Data, Decisions and Systems S.A.

jmartinez@daedalus.es

³ Universidad Autónoma de Madrid

{ana,sandoval}@maria.111f.uam.es

Abstract. This paper describes the adaptations of the MIRACLE group QA system in order to participate in the Spanish monolingual question answering task at QA@CLEF 2007. A system, initially developed for the EFE collection, was reused for Wikipedia. Answers from both collections were combined using temporal information extracted from questions and collections. Reusing the EFE subsystem has proven not feasible, and questions with answers only in Wikipedia have obtained low accuracy. Besides, a co-reference module based on heuristics was introduced for processing topic-related questions. This module achieves good coverage in different situations but it is hindered by the moderate accuracy of the base system and the chaining of incorrect answers.

1 Introduction

MIRACLE team submitted a run for the Spanish monolingual QA subtask at CLEF [4] that included as innovations: Wikipedia as an additional collection and the move towards topic related questions. Our basic QA system uses a pipeline architecture [3] and is based on Information Extraction. Most successful systems for Spanish have opted either for a similar strategy like Priberam [1] or text-mining like INAOE [5]. Our aim was to test the adaption of the QA system to other collections. Therefore we reused the basic QA system and developed a new module for merging answers based on temporal information. Finally, to cope with topic-related questions we developed linguistically motivated heuristics to identify the focus of a question and test their accuracy.

The rest of the paper is structured as follows, the next section describes the system architecture focusing on the new modules. Section 3 introduces the results

^{*} This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and projects by the Spanish Ministry of Education and Science (TIN2004/07083, TIN2004-07588-C03-02, TIN2007-67407-C03-01).

and a preliminary analysis of errors. Conclusions and directions for future work are discussed in Section 4.

2 System Overview

The architecture of the system (Figure 1) is similar to the multi-source architecture proposed in [2]. It is composed of two streams, one for each of the collections; EFE or Wikipedia. Each stream produces a ranked list of answers that are merged and combined by the Answer Source Mixer, a new module included for this evaluation. Question Analysis includes a new shared module for managing topic identification, context and anaphora resolution in topic-related question series. The basic system has been described in earlier participations [3] and it performs two kinds of operations; offline operations like indexing and document analysis and online operations like question analysis, sentence retrieval and answer selection.

2.1 Topic Identification in Topic-Related Questions

Introducing topic related questions requires a method to solve referential expressions that appear between questions and answers in the same question group. The system processes the first question and generates a set of candidates including the topic, the focus and the expected answer. A few rules that cover the most common cases are implemented to select the best topic for the question group. Rules use information available through question analysis and simplified assumptions about the syntactic structure of the questions.

The rules to locate the topic for a question group are :

- Answers of NUMEX subtype (numbers and quantities) are ignored as topics for questions series. The topic of the question, usually the syntactic subject will be the topic of following questions.
- Questions asking for a definition like *¿Quién es George Bush?* will add the topic and the answer (*presidente de los Estados Unidos*) to the group topic. An analog case occurs when we have questions like *¿Quién es el presidente de los Estados Unidos?*.
- Questions following the pattern *¿Qué NP * ?* " like *¿Qué organización se fundó en 1995?*. In these cases the noun group following the interrogative article is the focus of the question. Both the answer and the focus are added to the group topic.
- For the rest of the cases we use the answer as the topic.

Once the topic for the group is identified, the rest of the questions use it as an additional relevant term in order to locate documents and filter relevant sentences. Obviously, there is a problem when the system is not able to find the right answer and this is the topic for the rest of the group.

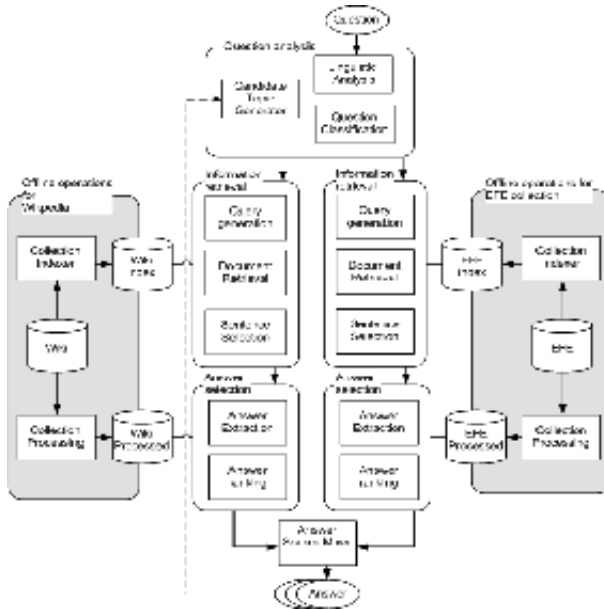


Fig. 1. MIRACLE 2007 System architecture

2.2 Combining EFE and Wikipedia Answers

The role of the Answer Source Mixer consists on combining answer candidates from the two different collections. We opted for a kind of ‘semantic’ reranking that takes into account the time period of collections, the verb tense and the time restrictions of the questions. In this way, no answer is really dropped from the candidates list but the list is reordered according to these clues. The heuristics used are the following:

- If the verb of the question appears in present tense preference is given to answers appearing in the Wikipedia collection.
- If the verb is in past tense and the question makes reference to the period covered by the EFE news collection, i.e., 1994 or 1995, then preference is given to answers from this collection.

3 Results and Error Analysis

Using the system described above, we have submitted one monolingual run for the Spanish subtask and obtained an overall accuracy of 15.00% (18.35% considering only factoids). Despite the inclusion of new sources of information like Wikipedia and the improvements carried in all preexisting modules, results obtained are lower than previous years [3]. However, this has been a general trend for all participants [4] due to the increasing difficulty of the task.

The complexity of the QA@CLEF task and the increasing number of stages that a question goes through in order to be answered, makes arduous any kind of error analysis. We reuse results presented in [4] and analyze them in the context of our system together with our own error analysis.

Regarding the main innovations, we have analyzed how the adaptations were able or not to solve them. There were 20 topic-related groups of questions with a total of 50 questions. The accuracy for the 170 questions that did not have to solve coreference was 15,29% while for the rest of the series is 13,33%. The difference is small and all additional errors except four are due to an incorrect selection of the first answer in the group.

In contrast, the analysis of the results for the different collections, EFE and Wikipedia, reveals that the source of the main decrease in accuracy is the strategy adopted for the latter one. For the 71 questions with answer in both collections the accuracy is 28,17%, slightly better than previous evaluations. When the answer could only be found in Wikipedia (114 questions), the accuracy decreases to 7.89%. This is specially accurate for definitional questions whose accuracy dropped to 3,13%. This reveals that the system strategies have been overadapted to the EFE collection over the years, for example with heuristics like pronominals for definitions.

4 Conclusions and Future Work

Result analysis shows that the source of most problems appear in the Wikipedia stream where we applied the same strategies used in EFE with little success. The module for coreference resolution is effective even if it uses few heuristics. In contrast, the greater contribution of errors is due to the low accuracy at finding the first answer. Alongside the improvement in general performance we plan to study methods to cope with several candidate answers and uncertainty when answering series of topic-related questions.

References

1. 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: *Evaluation of Multilingual and Multi-modal Information Retrieval*, pp. 300–309 (2007)
2. Chu-Carroll, J., Prager, J.M., Welty, C.A., Czuba, K., Ferrucci, D.A.: A Multi-Strategy and Multi-Source Approach to Question Answering. In: *TREC (2002)*
3. de Pablo-Sánchez, C., González-Ledesma, A., Moreno-Sandoval, A., Vicente-Díez, M.: MIRACLE experiments in QA@CLEF 2006 in Spanish: main task, real-time QA and exploratory QA using Wikipedia (WiQA) (2007)
4. Giampiccolo, D., et al.: Overview of the CLEF 2007 Multilingual Question Answering Track (2008)
5. Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., Montes-y-Gómez, M.: Using machine learning and text mining in question answering. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 415–423. Springer, Heidelberg (2007)

QA@L²F, First Steps at QA@CLEF

Ana Mendes, Luísa Coheur, Nuno J. Mamede, Ricardo Ribeiro,
Fernando Batista, and David Martins de Matos

L²F/INESC-ID Lisboa
Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
`qa-clef@l2f.inesc-id.pt`
`http://www.l2f.inesc-id.pt`

Abstract. This paper presents QA@L²F, the question-answering system developed at L²F, INESC-ID. QA@L²F follows different strategies according with the question type, and relies strongly on named entity recognition and on the pre-detection of linguistic patterns. Each question type is mapped into a single strategy; however, if no answer is found, the system proceeds and tries to find an answer using one of the other strategies.

1 Introduction

In this paper we present QA@L²F, the question-answering system from L²F, INESC-ID, as well as the results obtained at CLEF 2007.

In general terms, we can say that QA@L²F executes the following tasks:

- Information Extraction: information sources are processed, in order to extract potentially relevant information (such as named entities or relations between concepts), which is stored into a database;
- Question Interpretation: question is interpreted and mapped into an SQL query;
- Answer Finding: according with the question type, different strategies are followed in order to find the answer.

Considering information extraction, if a QA system focus on a particular domain or if the system is going to be used in an evaluation where the information sources are known, it makes sense to process all that information off-line, in order to get potentially relevant information. Thus, for the CLEF competition, QA@L²F pre-process the available corpora and gets structured information (such as named entities or noun phrases) that might be the answer to potential questions. This task is performed by many systems, as for instance Senso [14, 17].

Either to extract information or to interpret the question, some systems use natural language processing techniques [5, 8]; some perform named entity recognition and co-reference resolution. Also, many systems profit from thesaurus [2, 5, 8, 9] or ontologies [14, 17]. Internet may also be used as a resource [4, 6].

In what concerns QA@L²F, it profits from a Natural Language Processing (NLP) chain, which performs morpho-syntactic analysis, named entity recognition and shallow semantic analysis based on the named entities [10, 16]. This NLP chain uses the following tools:

- Palavroso [11], responsible for the morphological analysis and MARv [15] for its disambiguation;
- Rudrico (an improved version of PAsMo [13]), which not only recognize multi-word terms and collapse them into single tokens, but also splits tokens;
- XIP [1], which returns the input organized in chunks and connected by dependency relations.

This chain is used both in the information extraction step and in question interpretation.

In order to find the answer, systems such as INAOE [7] focus on the question type and follow different strategies according to it. QA@L²F also applies different strategies depending on the question type. However, if no answer is found, the system relaxes and tries to find an answer using one of the other strategies. Typically, several snippets are answer candidates and the QA system has to choose one of them. Although there are systems such as QUASAR [3] that combine frequency and the confidence given to both answer candidate and text passage in which the answer can be found, many systems choose the most frequent of all possible answers [18]. A confidence level is used by QA@L²F in one of its strategies; all the others only take frequency into consideration.

This paper is organized as follows: section 2 focus on the information extraction step; section 3 details the question interpretation; section 4 describes the different methods used to find the answer; section 5 presents and discusses the evaluation results; finally, section 6 concludes and points to future work.

2 Information Extraction

In order to extract information from newspaper corpora, a morpho-syntactic analysis is used to identify named entities, such as PEOPLE, which refer to person's names, CULTURE, to pieces of art, and TITLE, to person's professions and titles. With this information, as well as with a set of manually built linguistic patterns, relations between concepts are captured by the same NLP-chain, and stored into a database (from now on, the "relation-concepts" database). Every named entity recognized is also stored into a database (from now on, the "named entities" database).¹

For instance, consider the sentence "*Land and Freedom, de Ken Loach, evocação da Guerra Civil Espanhola*" ("*Land and Freedom, by Ken Loach, an evocation of the Spanish Civil War*"). In this piece of information might lay the answer to the question "*Who directed Land and Freedom?*". Therefore, by using linguistic patterns, the entry in the relation-concepts database from table 1 is built.

¹ As it should be clear, in both situations, the reference to the text snippet holding those relations/entities is also kept.

Table 1. Entry representing the relation between *Ken Loach* and *Land and Freedom*

CULTURE				
id	culture	author	confidence	count
1	Land and Freedom	Ken Loach	99	4

It should be noticed that these relation-concepts tables have information concerning the confidence given to that relation. It depends on the confidence level given to the linguistic patterns, which are assigned manually. Notice also, that “count” represents the frequency of this relation in the processed corpus.

In what concerns Wikipedia, QA@L²F used the WikiXML collection provided by the Information and Language Processing Systems group at the Informatics Institute, University of Amsterdam, as well as its database structure². A new table containing only the XML article nodes from every Wikipedia page, with no linguistic processing, was also created. The aim was to answer definition questions.

3 Question Interpretation

In QA@L²F, the question interpretation step is responsible for the transformation of the question into a SQL query.

The question is processed by: a) the NLP chain, which recovers the type of the question, as well as other information considered relevant (such as named entities and the question focus); b) a SQL generator.

Considering the question “*Quem é Boaventura Kloppenburg?*” (“*Who is Boaventura Kloppenburg?*”), after the NLP chain, both the type (WHO_PEOPLE) and the focus (*Boaventura Kloppenburg*) of the question are identified:

```
<DEPENDENCY name="WHO_PEOPLE">
<PARAMETER ind="0" num="11" word="Boaventura Kloppenburg"/>
</DEPENDENCY>
```

The SQL generator comprises the steps shown in Figure 1.



Fig. 1. SQL generation

The frame builder is responsible for choosing:

- the answer extraction script to be called next (depending on the type of the question);
- the question focus;
- all the named entities identified in the question.

² <http://ilps.science.uva.nl/WikiXML/>

The SQL generation is performed by a set of scripts that maps the frames into a SQL query.

Considering the previous example, the following frame is built:

```
SCRIPT    script-who-people.pl
TARGET    "Boaventura Kloppenburg"
ENTITIES  "Boaventura Kloppenburg" PEOPLE
```

This frame is then mapped into the following MySQL query, that will *possibly* retrieve the question's answer:

```
SELECT    title, confidence, count
FROM      FACT_PEOPLE
WHERE     name="Boaventura Kloppenburg"
GROUP BY  confidence DESC, count DESC
```

The “relation-concepts” database is queried and every title (or profession) connected with *Boaventura Kloppenburg* is retrieved, in descendant order of confidence and frequency.

4 Answer Finding

QA@L²F has a set of answer finding strategies. From within this set, the system has a preferred one to be applied on each question, depending on its type. The system expects this strategy to give the correct answer.

As an example, if the submitted question can be answered directly using the “relation-concepts” database, the system will just query that database. If not, the system adopts the following strategies, depending on the type of the question:

- *Linguistic Reordering*: the answer is searched in the wikipedia, after a re-ordering of some question elements;
- *Named Entities Matching*: the answer is searched in the named entities database;
- *Brute Force plus NLP*: some text snippets are chosen and processed in runtime; the obtained information provides QA@L²F a last chance to find an answer.

After detecting the type of question, one of these strategies is followed. If no answer is found, the system tries to answer it by using other strategy.

Using a method that allows it to jump to another strategy if the first one applied did not succeed, implicitly makes the system relax its constraints: it applies a strategy, even if it is not the one in which it relies the most to use on that question.

4.1 Linguistic Reordering

This strategy is used mainly for answering definition questions, like *Quem foi Pirro?* (*Who was Pirro?*) and *O que é a Igreja Maronita?* (*What is the Maronite*

Church?), or list questions, like *Diga uma escritora sarda.* (*Mention a sardinian writer.*).

QA@L²F uses Wikipedia in order to answer that group of questions. Firstly, the question interpretation step recovers the question focus (*Pirro, Igreja Maronita* and *escritora sarda*, considering the above examples). Then, it performs a search over the articles and applies the patterns inferred by the question structure to find the answer.

For definition questions, patterns are of the form: *question focus* plus the inflected verb *to be*. For instance, *Pirro foi...* (*Pirro was...*) or *Maronite Church é...* (*Maronite Church was...*). On the other hand, for list questions, those patterns are of the form: the inflected *to be* plus the *question focus*. For instance, *...é uma escritora sarda(...is a sardinian writer).*

This strategy is also used on questions for which the system could not find an answer using the linguistic patterns matching technique. Consider, for instance, the question *Quem foi Ésquilo?* (*“Who was Aeschylus?”*). The relation between *Ésquilo* and his title was not captured using linguistic patterns. Thus, the system searched on Wikipedia for the page having *Ésquilo* as title. The information about *Ésquilo*’s definition, a tragic greek poet, was found by processing the first line of this Wikipedia article page and, finally, returned as the question’s answer.

4.2 Named Entities Matching

This method queries the named entities database. A set of text snippets containing the named entities of the question is retrieved.

For instance, during the question interpretation of *Quem sucedeu a Augusto?* (*Who came after Augustus?*), the following frame was built:

```
TARGET EMPTY
ENTIDADES "Augusto " PEOPLE
AUXILIARES "sucedeu" ACTION "a Augusto"
```

With this information, QA@L²F searches on the database for snippets containing the named entity of type PEOPLE *Augusto* and the words *sucedeu* and *a Augusto*. For these last two, since they are not classified as named entities, the system performs a full-text query against the text snippets. The system gathers all the named-entities of types PEOPLE and PROPER (NAME) on those snippets, classifies them by order of frequency and returns the most frequent. Due to the fact that the system discards every candidate answer matching any word in the built frame, the named-entity *Augusto* is not chosen as the final answer.

4.3 Brute-Force Plus NLP

If none of the previously described strategies finds an answer, the system performs a full-text query against the raw text snippets database, returning the top ten best qualified snippets. Those snippets are processed by the NLP chain and the most frequent concept matching the wanted answer type is returned.

It should be noticed that this strategy is also used because we did not apply the information extraction module over the entire corpora. As so, although all the information is in the database, sometimes it is just in the form of a text snippet, without any processing. This technique allow us to extract information in run-time from paragraphs considered relevant.

4.4 Choosing the Answer

The system uses two main approaches in order to retrieve the final answer, depending on the strategy followed.

If the chosen strategy is either the linguistic patterns matching or the linguistic reordering, the system simply returns the answers found and takes in consideration the confidence and count attributes of each table (if they exist).

On the other hand, if the chosen strategy is either the named-entity recognition or the brute-force plus NLP, the answer extraction step depends on the type of the question. Having in mind that we are dealing with large corpora (564MB of newspaper text, both in European Portuguese and Brazilian Portuguese, as well as the Wikipedia pages found in the version of November, 2006), the system assumes that the correct answer is repeated on more than one text snippet. With this assumption, QA@L²F returns the most frequent named entity that matches the type of the question.

5 Evaluation

QA@L²F participated and was evaluated at CLEF for Portuguese as the query and target language. Table 2 presents the obtained results.

Table 2. QA@L²F results at CLEF 2007

Right	Wrong	ineXact	Unsupported	Total	Accuracy (%)
28	166	4	2	200	28/200 = 14%

Considering the correct answers:

- 11 were NIL;
- 3 followed the direct query of the “relation-concepts” database;
- 14 followed the linguistic reordering;
- from these 17, 2 used the relaxing mechanism.

It should be noticed that only 114 questions were interpreted (anaphora, ellipsis and some question types were not addressed).

Considering the ineXact answers, QA@L²F answered only the identified named entity, resulting into a ineXact answer. Nevertheless, it is difficult to be objective in deciding what should be the exact answer.

For instance, in the question “*Quem é George Vassiliou?*” (“Who is George Vassiliou?”) it is obvious that the answer “*presidente de Chipre*” (“Cypriot president”) is incomplete, as he was “*presidente de Chipre entre 88 e 93*” (“Cypriot

president between 88 and 93"). However, being given the following paragraph – “...norueguês, Henrik Ibsen, dramaturgo que escreveu *Peer Gynt*.” (“...norwegian, Henrik Ibsen, dramaturge that wrote *Peer Gynt*”) – it is not so obvious what should be the right answer to “*Quem foi Henrik Ibsen?*” (“Who was Henrik Ibsen?””).

If “*dramaturgo*” is incomplete, is “*dramaturgo norueguês*” enough? Or the right answer should be “*dramaturgo norueguês que escreveu Peer Gynt*”? It is difficult to decide.

Details about the evaluation can be found in [12].

6 Conclusions and Future Work

This paper presents QA@L²F first steps. The system follows different strategies according to the type of the submitted question and bases its performance on named entity recognition; if no answer is found, the system relaxes and tries to find the answer using another strategy.

Many improvements are yet to be done to QA@L²F. The improvement of all of the steps/techniques described in this paper are already scheduled, however the introduction of new strategies is also considered a goal.

Besides the current existence of a linguistic patterns matching approach, we would like to explore a syntactic pattern matching strategy, using patterns at the syntactic level.

We also would like to explore in detail Wikipedia’s standard structure (namely how it stores birth and death days and places, for instance), as it allows an easy retrieval of miscellaneous information.

References

1. A’it-Mokhtar, S., Chanod, J.-P., Roux, C.: A multi-input dependency parser. In: Proceedings of the Seventh IWPT (International Workshop on Parsing Technologies), Beijing, China (October 2001)
2. Amaral, C., Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., Vidal, D.: Priberam’s question answering system in qa@clef 2007. In: Working Notes for the CLEF 2007 Workshop (2007)
3. Buscaldi, D., Benajiba, Y., Rosso, P., Sanchis, E.: The UPV at QA@CLEF 2007. In: Working Notes for the CLEF 2007 Workshop (2007)
4. Cabral, L.M., Costa, L.F., Santos, D.: Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach. In: Working Notes for the CLEF 2007 Workshop (2007)
5. 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: Working Notes for the CLEF 2006 Workshop (2006)
6. Costa, L.: Esfinge - a modular question answering system for portuguese. In: Working Notes for the CLEF 2006 Workshop (2006)
7. Juárez-Gonzalez, A., Téllez-Valero, A., Denicia-Carral, C., Gómez, M.M.y., nor Pineda, L.V.: INAOE at CLEF 2006: Experiments in Spanish Question Answering. In: Working Notes for the CLEF 2006 Workshop (2006)

8. Laurent, D., Séguéla, P., Négre, S.: Cross Lingual Question Answer using QRISTAL for CLEF 2006. In: Working Notes for the CLEF 2006 Workshop (2006)
9. Laurent, D., Séguéla, P., Négre, S.: Cross Lingual Question Answering using QRISTAL for CLEF 2007. In: Working Notes for the CLEF 2007 Workshop (2007)
10. Loureiro, J.: NER - Reconhecimento de Pessoas, Organizações e Tempo. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal (2007)
11. Medeiros, J.C.: Análise morfológica e correcção ortográfica do português. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal (1995)
12. Mendes, A.: Clefomania, QA@L²F: Primeiros Passos. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal (2007)
13. Pardal, J.P., Mamede, N.J.: Terms spotting with linguistics and statistics. In: Proceedings of the international workshop Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués, IX Iberoamerican Conference on Artificial Intelligence (IBERAMIA 2004), November 2004, pp. 298–304 (2004)
14. Quaresma, P., Rodrigues, I.: A logic programming based approach to the QA@CLEF05 track. In: Working Notes for the CLEF 2005 Workshop (2005)
15. Ribeiro, R., Mamede, N.J., Trancoso, I.: Using Morphosyntactic Information in TTS Systems: comparing strategies for European Portuguese. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.d.G.V. (eds.) PROPOR 2003. LNCS, vol. 2721. Springer, Heidelberg (2003)
16. Romão, L.: NER - Reconhecimento de Locais e Eventos. Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal (2007)
17. Saias, J., Quaresma, P.: The Senso Question Answering Approach to Portuguese QA@CLEF-2007. In: Working Notes for the CLEF 2007 Workshop (2007)
18. Sarmiento, L.: Hunting answers with RAPOSA (FOX). In: Working Notes for the CLEF 2006 Workshop (2006)

Priberam's Question Answering System in QA@CLEF 2007

Carlos Amaral, Adán Cassan, Helena Figueira, André Martins,
Afonso Mendes, Pedro Mendes, Cláudia Pinto, and Daniel Vidal

Priberam Informática
Alameda D. Afonso Henriques, 41 - 2.º Esq.
1000-123 Lisboa, Portugal
{cma,ach,hgf,atm,amm,prm,cp,dpv}@priberam.pt

Abstract. This paper accounts for Priberam's participation in the monolingual question answering (QA) track of CLEF 2007. In previous participations, Priberam's QA system obtained encouraging results both in monolingual and cross-language tasks. This year we endowed the system with syntactical processing, in order to capture the syntactic structure of the question. The main goal was to obtain a more tuned question categorisation and consequently a more precise answer extraction. Besides this, we provided our system with the ability to handle topic-related questions and to use encyclopaedic sources like Wikipedia. The paper provides a description of the improvements made in the system, followed by the discussion of the results obtained in Portuguese and Spanish monolingual runs.

1 Introduction

Priberam has participated in the CLEF campaigns since 2005, where its QA system was evaluated in both monolingual [1,2] and cross-language [3,4] environments with rewarding results. In the framework of M-CAST¹, Priberam's system was also used for Portuguese, Polish and Czech and applied to a digital libraries project. For QA@CLEF 2007, we focused our participation on the Portuguese and Spanish monolingual tasks.

This year, the CLEF QA track presented two novelties with direct consequences in the evaluation of QA systems. First, the organisation introduced topic-related questions, that is, questions clustered around a common topic that might present anaphoric links between them. Second, it also added the open domain encyclopaedia Wikipedia as a target document collection to the already existent newspaper corpora. Although its overall architecture was maintained,

¹ M-CAST—Multilingual Content Aggregation System based on TRUST Search Engine—was an European Commission co-financed project (EDC 22249 M-CAST), whose aim was the development of a multilingual infrastructure enabling content producers to access, search and integrate the assets of large multilingual text (and multimedia) collections (<http://www.m-cast.infovide.pl>).

several changes were made to the Priberam's QA system, both in the Portuguese and in the Spanish modules, the most relevant one being the introduction of syntactical question processing. These modifications, together with the introduction of the two simultaneous changes by the CLEF organisation, had an impact in the performance of our system, as we will show in the analysis of the results.

This paper is organised as follows: in Sect. 2, we present the major improvements made to the system since last year's QA@CLEF edition; in Sect. 3 we analyse and discuss the results of both monolingual runs; finally in Sect. 4 we present the conclusions and future work.

2 Adaptations and Improvements of the System

In this section, we present the major adjustments made to our system to meet the new challenges proposed in this year's QA@CLEF: the work done in the syntactical processing of the question, and the necessary adaptations to deal with topic-related questions and with the Wikipedia encyclopaedic source. We also mention other linguistic tools and resources that were developed.

2.1 Syntactic Analysis of Questions

As reported in [2,4], Priberam's QA system is based upon a five-step architecture: the indexing process, the question analysis, the document retrieval, the sentence retrieval, and the answer extraction. When a question is submitted and matches a given question pattern (QP), a category is assigned to it and a set of question answering patterns (QAPs) becomes active. Then, documents containing sentences with categories in common with the question (previously determined during indexation via answer patterns (APs)) are analysed; the active QAPs are then applied to each sentence in order to extract the possible answers. While the overall architecture remains unchanged, this year, following the conclusions taken from preceding evaluations, we implemented a mechanism for the syntactic treatment of questions.

In the former version of the QA system, the question analysis stage categorised questions according to a previously defined question typology, by matching a set of question patterns [2]. More than one category per question was allowed. Although a multicategorisation scheme has the advantage of allowing more than one category in cases where it is difficult to choose only one, the excess of categories was one of the causes for errors in the extraction of candidate answers in our former CLEF participations. In the present version, and taking advantage of the company's linguistic technology developed for FLiP,² QPs were enhanced

² *Ferramentas para a Língua Portuguesa*, Priberam's proofing tools package for Portuguese. FLiP includes a grammar checker, a spell checker, a thesaurus and a hyphenator that enable different proofing levels—word, sentence, paragraph and text—of European and Brazilian Portuguese. An online version is available at <http://www.flip.pt/online>.

with syntactical information. When a question is submitted, its syntactic structure is captured in a parsing stage that also determines the syntactic function of the question pivots. Parsing the question allows to determine its main syntactic constituent (the *object*), as well as its secondary syntactic constituents, in case they exist. For instance, in question 8 of the Portuguese test set “De que estado brasileiro foi governador Adhemar de Barros?” [Of which Brazilian state was Adhemar de Barros governor?], the object is *Adhemar de Barros* and the secondary constituents are *governador* and *estado brasileiro*. This information is then used to validate the category assigned to each question and to allow a more exact, syntactically based answer extraction. By differentiating pivots through syntactic tags, we also improve the document retrieval stage, especially when it comes to long questions where there are too many pivots. In the answer extraction stage, we enhanced QAPs to include information about each pivot’s syntactic specifications, which the system tries to match with the answer pivots. The syntactic information captured by parsing may also be useful for answer validation, an aspect that we intend to exploit in the future.

2.2 Handling Topic-Related Questions

Priberam’s QA system was adapted to handle topic-related questions; these are clusters of questions related to the same topic and possibly containing anaphoric references between one question and the others. Although questions belonging to the same cluster are identified as such, no information is given about the topic, which has to be inferred from the first question/answer pair.

To emulate an interactive task, our approach requires only one question to be processed at a time, which means that the topic detection is performed regardless of any information contained in the subsequent questions. This requirement makes one-topic based approaches (those in which only one topic is considered) unsuitable, because one cannot infer from the first question which will be the topic of the whole cluster. For example, consider the question 58 of the Spanish test set, “¿Quién diseñó el procesador Zilog Z80?” [Who designed the Zilog Z80 processor?], whose answer is “Federico Faggin.” Only after parsing the subsequent question, “¿De cuántos bits era este procesador?” [How many bits was this processor?] could the system have inferred that the topic was “Zilog Z80,” and not “Federico Faggin.” To overcome this difficulty, we designed the system to accept *several* topics, through the following procedure (illustrated in Table 1):

1. The system parses the first question of a cluster and follows the usual procedure to extract an answer (if any is available);
2. The system calls a method `GetTopic` that collects the noun phrases (nouns, proper nouns and named entities) of the extracted answer and the noun phrases of the question which were considered object pivots (see Subsect. 2.1), merging these two groups into a single list of *topic pivots*;
3. In subsequent questions from the same cluster, the system first parses the question, obtaining its pivots, and then calls a method `SetTopic`, which appends to the question’s pivots the list of topic pivots collected in the previous step.

Table 1. Procedure for topic-related questions

Question	“¿Quién diseñó el procesador Zilog Z80?”
Pivots	<i>diseñar, procesador, zilog z80</i>
Extracted Answer	<i>Federico Faggin</i>
Topic pivots after GetTopic	<i>federico faggin, zilog z80</i>
Question	“¿De cuántos bits era este procesador?”
Pivots after SetTopic	<i>bits, procesador</i> + <i>federico faggin, zilog z80</i>
Extracted Answer	<i>8</i>

We are considering as future work a more sophisticated approach that involves anaphora resolution and through which we can do some sort of topic disambiguation by choosing only the co-references (i.e. the topics) that best suit the question. In the above example, the expression “este procesador” could only refer back to “Zilog Z80” and not to “Federico Faggin”; hence the topic pivot “Federico Faggin” could have been discarded if an anaphora resolution method had been used to disambiguate the topic.

2.3 Addition of the Wikipedia Collection

Unlike the collections of newspaper articles, the Wikipedia collection has a rich structure (links, categories, disambiguation pages, etc.) that suggests using strategies capable of extracting knowledge from structured data. However, as a first approach, we simply indexed the Wikipedia articles as natural language text, with some minor adaptations. The indexation module was designed to ignore all the metadata included in tables and boxes, all disambiguation and discussion pages, and any internal Wikipedia pages whose title starts with “Wikipedia:”. Links of the form “[[<Article title>|<Link text>]]” were converted into strings like “<Link text> (<Article title>)” (i.e., putting the title of the linked article between parentheses) and indexed as natural text. This strategy allows answering short definition questions (e.g. acronyms) without making any change in the system modules. We developed a simple scheme of anaphora resolution for Wikipedia articles: since many anaphoric references have the article title as their referent, every time we parse a sentence and find a null subject or a personal pronoun subject, we replace it by the article title and parse the sentence again.³ Consider the question 136 of the Spanish test set “¿Cuánto mide de alto la Pirámide del Sol de Teotihuacan?” [How high is the Pyramid of the Sun at Teotihuacan?]. The following sentence appears in the article titled “Pirámide del Sol (Teotihuacan)” and the parser detects that it has a null subject: “Tiene 65 m de altura.” So a new sentence is composed, “Pirámide del Sol (Teotihuacan) tiene 65 m de altura,” and the answer *65 m* is successfully extracted.

³ Of course, this approach excludes other possible referents besides that expressed in the article title, which is a limitation.

2.4 Changes in the Processing Modules and Resources

Some changes were also made in the processing modules, specifically with respect to the way QPs, APs and QAPs are parsed (see [2,4] for more details). We have adapted Earley’s parsing algorithm [5] to be able to handle our grammar for QA. This allowed us to introduce some new rules that take profit of grammar recursion. For instance, a new **Rep** command was introduced to deal with the arbitrary repetition of a term (*e.g.*, **Rep** [Cat(N) Cat(Vg)] stands for an arbitrary sequence of nouns followed by a comma); this feature provides an efficient method to extract answers for list questions. Another feature that uses recursion is the ability to follow different paths to ignore or to take into account text between parentheses, when testing a pattern. For example, consider the sentence “Jorge Sampaio (presidente de Portugal) deslocou-se em visita de estado à República Popular da China” [Jorge Sampaio (president of Portugal) has paid a state visit to the People’s Republic of China]. The text between parentheses could be extracted as an answer to the question “Quem é Jorge Sampaio?” [Who is Jorge Sampaio?]. But if the question is “Que país visitou Jorge Sampaio?” [Which country did Jorge Sampaio visit?], it would be useful to ignore this portion of text. Using our adaptation of Earley’s parser, both paths are explored when searching for an answer. This turns out to be particularly useful for Wikipedia articles, since we convert links to other articles into the article titles between parentheses (see Subsect. 2.3).

The Spanish language resources were also enhanced with the improvement of the lexicon, the recognition of named entities and the inclusion of a Spanish thesaurus.

3 Results and Discussion

We now present the results and discuss the validity of our choices. The sets of questions were classified according to three question categories: *factoid* (FACT), *definition* (DEF) and *list* (LIST). The official assessments for both tracks and a comparative analysis with other systems is provided in [6]; for the monolingual Portuguese (PT) task, we reproduce the official assessments in Table 2; for Spanish (ES), we present instead our internal evaluation, as we find it more useful for the discussion that follows, as it reflects better the strengths and weaknesses of the system. We refer to [7] for further discussion regarding the differences between our evaluation and the official results.

The general results of Table 2 show a significant decrease of overall accuracy in PT and an increase in ES when comparing with the last CLEF campaign [4]. However, analysing separately the accuracy of non-topic-related questions (those questions that could be directly answered without any topic detection) we can see that the performance in PT was quite satisfactory. There are some noticeable differences between the number and size of the question clusters in PT and ES: while PT has 25 clusters including 75 questions (with many clusters having 4 questions), ES has 20 clusters including 50 questions (mostly with 2 questions per cluster). Therefore, this task was far more difficult for PT, which could have led

to its lower results. When comparing with last year's result, we should also take into account that this year's inclusion of the Wikipedia collection increased the level of difficulty. The greater volume of information from the combined corpora was in some cases a handicap to reach the best answers. Syntactic processing helped us, though, to better tune the categorisation of the questions and to structure the information provided by the question pivots; and the improvements in the Spanish modules raised the accuracy of the ES run.

Table 3 displays the distribution of errors along the main stages of Priberam's QA system. In the ES run, the reasons for wrong answers are fairly the same as in last year, with most errors occurring during the extraction of candidate answers. In the PT run, however, there are now many errors occurring during the document retrieval stage, especially in the topic-related questions. A possible explanation is that the topic may be erroneously detected, resulting in too many pivots being assigned to the subsequent questions, which causes the system to

Table 2. Results by category of question, including detailed results of topic and non topic-related questions

		R		W		X		U		Total		Accuracy (%)	
		PT	ES	PT	ES	PT	ES	PT	ES	PT	ES	PT	ES
Non-topic related	FACT	63	77	43	57	3	2	1	1	110	137	57.3	56.2
	DEF	25	19	2	7	5	0	0	0	32	26	78.1	73.1
	LIST	4	2	4	5	0	0	0	0	8	7	50.0	28.6
	Total	92	98	49	69	8	2	1	1	150	170	61.3	57.6
Topic related	FACT	8	11	39	16	2	1	0	0	49	28	16.3	39.3
	DEF	0	0	0	0	0	0	0	0	0	0	-	-
	LIST	0	0	1	2	0	0	0	0	1	2	-	0.0
	Total	8	11	40	18	2	1	0	0	50	30	16.0	36.7
General (all)	FACT	71	88	82	73	5	3	1	1	159	165	44.7	53.3
	DEF	25	19	2	7	5	0	0	0	32	26	78.1	73.1
	LIST	4	2	6	7	0	0	0	0	9	9	44.4	22.2
	Total	100	109	90	87	10	3	1	1	200	200	50.0	54.5

Table 3. Reasons for W, X and U answers

Stage ↓	Question →	W+X+U		Failure (%)	
		PT	ES	PT	ES
Document retrieval		45	16	45.0	17.6
Extraction of candidate answers		23	37	23.0	40.7
Choice of the final answer		21	29	21.0	31.9
NIL validation		4	6	4.0	6.6
Other		7	3	7.0	3.3
Total		100	91	100.0	100.0

Table 4. Example of a double-topic cluster of questions. Notice the change of topic from “Bill Gates” (question ids 185 and 186 of the ES set) to “Universidad de Harvard” (question ids 187 and 188).

Id	Group id	Question
185	2161	¿Cómo se llama la mujer de Bill Gates? [What is the name of Bill Gates’ wife?]
186	2161	¿En qué universidad estudiaba él cuando creó Microsoft? [At which university was he studying when he created Microsoft?]
187	2161	¿Qué presupuesto tenía esa universidad en 2005? [What was the budget for that university in 2005?]
188	2161	¿En que [sic] año se fundó? [In which year was it founded?]

miss the documents in which the correct pivots appear. Some of the wrong answers classified as *Other* are due to dubious clusters of questions that seem not to respect the guidelines, since topics do not always come from the first question/answer pair, as illustrated in Table 4.

There are some examples too, in both languages, of answers that apparently can only be extracted from tables or boxes, such as the coordinates of Guarda for the PT question 87 “Qual é a latitude e longitude da Guarda?” [What are the latitude and longitude of Guarda?]. In some cases, the documents were retrieved, but the answer was not extracted.

To sum up, this year’s CLEF campaign set forth new challenges to our QA system. Due to the two major changes (the introduction of topic-related questions and the addition of the Wikipedia collection), we cannot directly compare the accuracy of the Priberam’s QA system with previous years. By limiting the analysis to non-topic-related questions, we observe, however, a significant improvement in Spanish and a similar performance in Portuguese. Besides, our system achieved a more accurate question categorisation due to the introduction of syntactical parsing during question processing, decreasing the number of wrong candidate answers. Given the short amount of time available, we relied on simple approaches to handle the Wikipedia collection and to answer topic-related questions; we believe that more sophisticated strategies would lead to a much better performance. On the one hand, a deeper treatment of topic-related questions is necessary, with a more accurate topic identification and a strategy for co-reference resolution, since our simple approach often leads to questions with many pivots, some of them outliers. On the other hand, the absence of a specific information extraction technique specifically tailored for the Wikipedia collection (*e.g.*, to extract information from tables or lists) may also explain why some questions were not correctly answered. We also did not exploit the hyperlink structure of Wikipedia. All of these issues are subject of current work.

4 Conclusions and Future Work

We expect to further improve the syntactical analysis and to extend it to the answer extraction module. The answer syntactic analysis will allow the system to more precisely match the pivots of the question with their counterparts in the answer, taking into account their syntactic functions. A further development of anaphora resolution will also be one of our goals in the future. We expect to broaden the approach applied this year in Wikipedia article titles, by using our syntactic parsing engine to deal with co-references also in text sentences.

Finally, we intend to evaluate again the cross-language performance of our system, as this was left out in this campaign. In particular, it would be interesting to evaluate how the system performs with topic-related questions in a cross-language environment.

Acknowledgments

Priberam Informática would like to thank Synapse Développement, TiP, University of Economics of Prague (UEP), as well as the CLEF organisation and Linguatca. We would also like to acknowledge the support of the European Commission in the M-CAST (EDC 22249 M-CAST) project.

References

1. Vallin, A., 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: Working Notes for the CLEF 2005 Workshop, Vienna, Austria, 21-23 September (2005)
2. Amaral, C., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C.: Priberam's question answering system for Portuguese. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria, 21-23 September (2005)
3. 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 CLEF 2006 Workshop, Alicante, Spain, 20-22 September (2006)
4. 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: Working Notes for the CLEF 2006 Workshop, Alicante, Spain, 20-22 September (2006)
5. Earley, J.: An efficient context-free parsing algorithm. *Comm. ACM* 13, 94-102 (1970)
6. Giampiccolo, D., Peñas, A., Ayache, C., Cristea, D., Forner, P., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 multilingual question answering track. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 9-21 September (2007)
7. Amaral, C., Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., Vidal, D.: Priberam's question answering system in QA@CLEF 2007. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary, 9-21 September (2007)

Combining Logic and Aggregation for Answer Selection

Ingo Glöckner

Prakt. Informatik VII, University of Hagen, 58084 Hagen, Germany

Abstract. MAVE (Multinet-based Answer Verification) is a system for answer validation which combines logic-based techniques and aggregation for identifying the correct answers in given sets of answer candidates. The paper explains the basic concepts underlying MAVE and also presents ablation studies which reveal the contribution of the proposed methods to the achieved quality of selection.

1 Introduction

Answer selection is one of the key components of a QA system. The selection task can be described as follows: a) Start from a *validation set* of *validation items* (q, α, w) with question q , answer string α and supporting text passage or ‘witness text’ w . b) For *answer selection*, determine the ‘best’ validation item v and label v as SELECTED or REJECTED. c) For a complete *answer validation*, also mark the remaining items as VALIDATED or REJECTED. While machine learning and approaches to recognizing textual entailment are popular choices for answer validation – see [1] for an overview of the techniques used in the Answer Validation Exercise (AVE) 2007 – it is more typical of QA systems to exploit redundancy by aggregating evidence (e.g. by selecting the most frequent answers). It is not obvious how the logic-oriented techniques for answer validation and the redundancy-based techniques for selection must be combined in order to maximize selection quality. The paper describes a possible approach which proved effective in the AVE 2007. The contribution of the main techniques is also investigated.

2 System Description

This section explains the system architecture of MAVE (see [2] for details).

Deep linguistic analysis. WOCADI [3], a robust parser for German, is used for deep linguistic analysis of question, answer, and supporting text passage. This step also involves coreference resolution. The postprocessing of parsing results includes a *synonym normalization* by replacing all lexical concepts with canonical synset representatives. It is based on 48,991 synsets (synonym sets) for 111,436 lexical constants.

Hypothesis construction. Question and answer together express a *hypothesis* to be checked against the supporting text. Thus, ‘*At which age did Elvis die?*’ and

the answer ‘43’ constitute the hypothesis ‘*Elvis died at the age of 43.*’ Forming a textual hypothesis is hard for highly inflecting languages like German. MAVE avoids this problem by directly building a logical hypothesis from the logical analysis of question and answer.

Robust entailment test. The answer is logically validated by proving the hypothesis from the logical representation of the supporting text. Robustness against knowledge gaps and errors of semantic analysis is achieved by embedding the prover in a relaxation loop which repeatedly skips literals of the hypothesis until a proof of the remaining query succeeds. The skipped literal count is used as the entailment indicator. The prover utilizes 10,000 lexical-semantic facts and 109 implicative rules from AVE 2006.

Fallback entailment test. The logic-based criterion is only defined if question, answer and supporting text can be parsed; otherwise only simple lexical overlap is used as a replacement. This overlap-based fallback method uses synonym normalization and expands concepts by applying lexical-semantic relations (currently 27,814 nominalizations of verbs and 15,052 nominalizations of adjectives). Scope of the matching is restricted to sentences, with the best sentence determining the result of matching.

Assigning failure probabilities. Comparable scores are needed for ranking and aggregation, but the logic-based error counts and overlap-based counts (to which MAVE backs off when NL analysis fails) are incommensurable. The system thus switches from error counts to the probability that the answer is correct (or wrong) given the error count. The probabilities $errProb = P(\text{supported answer incorrect} | \text{error count of validation item} = k)$ for both types of error count are estimated from the AVE 2007 training set.

Aggregating evidence. MAVE aggregates the available evidence when several text passages support the same answer. Obviously an answer is logically justified if it is logically justified from at least one supporting text, i.e. $multErrProb = \prod_c errProb_c$ (assuming independence), where c ranges over all validation items supporting the same answer. This approach proved too optimistic when there is a lot of redundancy, though, so that MAVE now uses a more stable pragmatic criterion (with $minErrProb = \min_c errProb_c$), viz $combinedErrProb = (errProb + multErrProb + minErrProb)/3$.

Determining additional quality factors. The aggregated error probability coincides for all validation items supporting the same answer. A heuristic quality factor $wnHeuristicQual$ for preferences on ‘good’ supporting texts is used for tie-breaking [2]. The heuristic quality factor for answers, $awHeuristicQual$, also includes sanity criteria: a) a test for trivial answers entailed by the query: ‘*Who is Gianni Versace?*’ – ‘*Versace*’; b) a circularity test: ‘*Who is the inventor of the car?*’ – ‘*The inventor of the modern car*’; c) a test for other non-informative definitions, e.g. isolated nomina agentis or role terms: ‘*Who is Vitali Klitschko?*’ – ‘*The brother.*’ d) a check for mismatch of expected vs. actual answer type.

‘When did Google publish the Google Web API?’ – ‘a software’. All of these criteria do not depend on the supporting text and are thus not aggregable.

Total validation score. The final score used in selection and validation decisions is $validationScore = awHeuristicQual \cdot (wnQual + bonusWnQual) / 2$, where $wnQual = wnHeuristicQual \cdot (1 - combinedErrProb)$, and the term $bonusWnQual$ denotes the maximal $wnQual$ achieved by any validation item supporting the considered answer. The bonus term utilizes that there are typically few ‘un-supported’ answers compared to wrong ones. Thus correctness of an answer, as judged from the best-supporting snippet, also hints at the validity of other validation items for this answer.

Decision rules for selection and validation. MAVE uses separate thresholds for selection/rejection of the best answer and validation/rejection of remaining alternatives: a) SELECT the validation item with highest $validationScore$ in the test set if $validationScore \geq fSelThresh$, REJECT otherwise. b) VALIDATE the remaining validation items in the test set if $validationScore \geq fValThresh$, REJECT otherwise. Separate thresholds make sense because factual questions often have only one correct answer. This suggests a stricter criterion for alternatives. In order to extract thresholds for optimal f-measure, MAVE chooses $fSelThresh$ and $fValThresh$ with $fValThresh \geq fSelThresh$ such as to maximize f-measure over the training set. For optimal selection rate, MAVE always selects the best given answer ($fSelThresh = 0$), and the threshold $fValThresh$ for non-best answers is chosen to maximize f-measure on the training set.

Extending the scope of aggregation. Aggregation of evidence needs not be restricted to validation items which support identical answers. The *cluster method* applies a simplification function σ to the answer strings which converts to lowercase, removes accents, eliminates stopwords, etc. Scope of aggregation for the considered answer α is then extended to all validation items supporting the same answer cluster, i.e. an answer α' with $\sigma(\alpha) = \sigma(\alpha')$. This method works well for answer variants, but there is no simple extension to inclusions. In this case aggregation based on containment of cluster keys $\sigma(\alpha) \sqsubseteq \sigma(\alpha')$ will sometimes be incorrect ($\alpha = \text{‘a prophet’}$ vs. $\alpha' = \text{‘a false prophet’}$). An alternative method called *evidence reassignment* (ERA) restructures the validation set by assigning each piece of evidence to all answers potentially supported by it. This process must be backed by methods for spotting answer variants and inclusions. Following the reassignment of supporting text passages, validation and aggregation of evidence is performed only for identical answers. In this way ERA copes with non-monotonic NL constructions. Consider the validation item (which should be rejected): $v = (\text{‘Who is Di Mambro?’}, \text{‘a prophet’}, w, 1)$, where $w = \text{‘... self-proclaimed prophet Di Mambro...’}$, and the last component $o = 1$ marks the item as non-generated. A simple method for spotting inclusions might wrongly propose $v' = (\text{‘Who is Di Mambro?’}, \text{‘a false prophet’}, w', 1)$, with $w' = \text{‘... Di Mambro, the false prophet...’}$, as including v . ERA avoids the false conclusion that Di Mambro is a prophet since it forms a new item

$v'' = ('Who\ is\ Di\ Mambro?', 'a\ prophet', w', 0)$ replacing v' . Thus a proof that Di Mambro is a prophet fails, and v is indeed rejected.

Active enhancement of validation sets. The lack of redundancy in the AVE 2007 test set suggested actively generating redundancy by adding more supporting text passages. Thus three QA systems were run on the AVE 2007 questions (see [2]). These runs produced 12,837 answer candidates with 30,432 supporting passages, which were then searched for inclusions with respect to the AVE 2007 answers. In this way, the original test set with 282 validation items was enhanced by 2,320 auxiliary validation items.

3 Evaluation

MAVE was evaluated on the AVE 2007 test set for German [1]; see Table 1 which also lists the reference results of the current version of MAVE. Here, CF means clustering of answers and optimizing thresholds for f-measure, CQ means clustering and optimizing for qa-accuracy, EF means ERA method and optimizing for f-measure, EQ means ERA optimizing for qa-accuracy, and * marks the current results of MAVE. after further debugging. A series of ablation results is shown in Table 2. Extracting additional supporting text passages had a positive effect: The PCF run (no enhancement, optimizing f-measure) loses 5% in f-measure compared to EF* and 17% in selection rate. The PCQ run

Table 1. AVE 2007 results of MAVE and reference results of current system. The metrics shown are f-measure, precision, recall, qa-accuracy (correct selection per question) and selection rate (correct selection per question with an answer in the test set).

model	f-meas prec		recall	qa-acc	sel-rate
CF (Run1)	0.72	0.61	0.90	0.48	0.89
CF*	0.73	0.62	0.90	0.49	0.90
CQ*	0.70	0.56	0.94	0.50	0.93
EF*	0.73	0.62	0.91	0.50	0.92
EQ (Run2)	0.68	0.54	0.94	0.50	0.93
EQ*	0.69	0.55	0.93	0.50	0.93

Table 2. Results without enhancing validation sets (PCF, PCQ), with joint thresholds for selection/validation (EJF, EJQ), without sanity checks (E-F, E-Q), without logical features (LEF, LCF, LEQ, LCQ) and without lexical-semantic relations (KCF, KCQ)

model	f-meas prec		recall	qa-acc	sel-rate	model	f-meas prec		recall	qa-acc	sel-rate
PCF	0.68	0.61	0.76	0.41	0.75	LEF	0.72	0.59	0.94	0.49	0.90
PCQ	0.66	0.53	0.88	0.48	0.89	LCF	0.72	0.59	0.93	0.48	0.89
EJF	0.68	0.55	0.88	0.46	0.85	KCF	0.56	0.44	0.78	0.42	0.77
EJQ	0.45	0.29	0.97	0.50	0.93	LEQ	0.68	0.53	0.96	0.50	0.92
E-F	0.68	0.55	0.90	0.48	0.89	LCQ	0.68	0.53	0.96	0.50	0.92
E-Q	0.65	0.51	0.91	0.49	0.90	KCQ	0.55	0.43	0.79	0.42	0.79

(plain validation sets, selection-oriented) also loses 3% of f-measure and 4% of selection rate compared to EQ*. The use of separate thresholds for selecting the best answer and for accepting alternatives is also effective: EJF (ERA with joint threshold, optimizing for f-measure) loses 5% of f-measure and 7% of selection rate compared to EF*. EJQ (ERA with joint threshold, optimizing for qa-accuracy) keeps a selection rate of 0.93, but suffers a loss of f-measure by 24% compared to EQ*. The sanity checks of MAVE show a positive contribution of 5% f-measure comparing E-F to EF* (or 4% comparing E-Q to EQ*). Table 2 also reveals a very small positive effect of logical inference and a strong positive effect of lexical-semantic knowledge. The success of the fallback method which lacks a structural comparison might be an artifact of the AVE 2007 setup: since the answers originate from state-of-the-art QA systems with their own tests for structural match, repeating such a test once again is no longer effective.

4 Conclusion and Future Work

The MAVE system demonstrates how answer selection rates can be boosted by combining techniques for answer validation and for leveraging redundancy. Current work aims at *real time answer validation*, i.e. managing system resources in such a way that the best answer is also found under pre-defined time constraints.

References

1. Peñas, A.: Rodrigo, I., Verdejo, F.: Overview of the answer validation exercise 2007. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
2. Glöckner, I.: University of Hagen at QA@CLEF 2007: Answer validation exercise. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
3. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück (2003)

On the Application of Lexical-Syntactic Knowledge to the Answer Validation Exercise*

Óscar Ferrández, Daniel Micol, Rafael Muñoz, and Manuel Palomar

Natural Language Processing and Information Systems Group
Department of Computing Languages and Systems
University of Alicante, Spain
{ofe,dmicol,rafael,mpalomar}@dlsi.ua.es

Abstract. This paper presents a system that applies Textual Entailment recognition techniques to the AVE task. This is performed comparing representations of text snippets by means of a variety of lexical measures and syntactic structures. The representations of the question and the answer are compared determining if there is an entailment relation between them. The performed experiments over the English test corpus obtained a maximum F-score of 0.39.

1 Introduction

With our participation in the *Answer Validation Exercise* (AVE), we would like to evaluate our system, which is designed to recognize Textual Entailment relations, within the realistic environment that this task provides [1]. In addition, this allows us to apply our system in the field of QA, which is very appealing. An application of our system within the *Third ACL-PASCAL Recognising Textual Entailment (RTE) Challenge* is described in [2]. To apply our system to the AVE competition we had to perform some preprocessing that will be explained in detail later.

2 System Core

The approach proposes several methods that mainly rely on lexical and syntactic inferences in order to address the recognition task and detect implications between two text snippets (the text or the passage and the hypothesis that will be generated by both the question and the answer).

Corpora Processing: since our system is designed to determine implications between two texts, the best way to adapt the AVE corpus to it was, for each answer and question, to convert them into an affirmative sentence and detect

* This research has been partially subsidized by the QALL-ME consortium, 6th Framework Research Programme of the European Union (EU), FP6-IST-033860. It has also been supported by the Spanish Government project CICYT TIN2006-1526-C06-01 and the Spanish Generalitat Valenciana project ACOM06/90 and ACOMP07/056.

if there is entailment with its associated text. A set of regular expressions were used to perform these conversions.

We attempted to match every English question-answer pair with the set of regular expressions. However, for the pair that did not match with any regular expression two approaches were developed: in the first one, called *automatic*, the words of the answer are grouped with the ones of the corresponding question, while for the second one, called *semi-automatic*, we have done a manual review of these pairs generating the affirmative sentences.

Lexical module: it relies on the computation of a wide variety of lexical measures, which basically consist of overlap metrics, and their integration within a machine learning algorithm. Some researchers have already used this kind of metrics, but in contrast to them our approach does not use semantic knowledge. Prior to the calculation of the measures, several data structures containing the stems, lemmas and functional words were generated in order to apply the measures over them. The considered lexical measures are:

- **Simple matching:** binary normalized matching between the data structures extracted from the two snippets.
- **Levenshtein distance:** it is similar to simple matching, but using the Levenshtein distance as matching measure.
- **Consecutive subsequence matching:** this measure assigns the highest relevance to the appearance of consecutive subsequences, from length two until the length in words of the hypothesis. It assigns the same relevance to all consecutive subsequences with the same length. Furthermore, the longer the subsequence is, the more relevant it will be considered.
- **Tri-grams:** binary normalized matching between tri-grams of letters belonging to both snippets.
- **ROUGE measures:** due to the fact that these measures are very related to word overlapping and considering the impact of n-gram overlap metrics in textual entailment, we decided to integrate them¹ in our system. We have implemented these measures as defined in [3].

Syntactic module: it aims to provide a good accuracy rate by using few syntactic modules that behave collaboratively:

- **Tree generation:** constructs the corresponding syntactic dependency trees. For this purpose, *MINIPAR* [4] output is generated and afterwards parsed for each text and hypothesis in our corpus.
- **Tree filtering:** discards irrelevant data. We only consider those words whose grammatical category belongs the most relevant ones².
- **Graph node matching:** in this stage we proceed to perform a graph node matching process, termed alignment, which consists in finding pairs of words in both trees whose lemmas are identical, no matter whether they are in the same position within the tree. Some authors have already designed similar matching techniques, although these include semantic constraints that we have decided not to consider.

¹ ROUGE-N (n=2 and n=3), ROUGE-L, ROUGE-W and ROUGE-S (s=2 and s=3).

² Details about these categories can be found in [5].

Now we will describe the similarity rate calculation based on syntactic knowledge. Let τ and λ represent the text's and hypothesis' syntactic dependency trees, respectively. We assume we have found a word, namely β , present in both τ and λ . Now let γ be the weight assigned to β 's grammatical category, σ the weight of β 's grammatical relationship³, μ an empirically calculated value that represents the weight difference between tree levels, and δ_β the depth of the node that contains the word β in λ . We define the function $\phi(\beta) = \gamma \cdot \sigma \cdot \mu^{-\delta_\beta}$ as the one that calculates the relevance of a word in our system. The experiments performed reveal that the optimal value for μ is 1.1.

3 Experiments and Results

Our system returns a numeric value to determine the entailment relation level. To decide which answer is marked as SELECTED, we chose the one with the highest entailment score among all answers to the same question. It is quite difficult to choose one of the VALIDATED values as SELECTED since differences between entailment scores are usually very low. This happens due to the fact that no semantic knowledge is considered. Therefore, although the system is able to determine lexical-syntactic implications, in the case of SELECTED values this does not seem to be enough. Table 1 shows the results obtained.

Table 1. Results obtained for the AVE 2007 English corpora

Corpus	Run	Prec.	Rec.	F-measure	Q-A acc.	AVE ranking
Dev	baseline ⁴	0.12	1.0	0.21	–	–
	lex automatic	0.26	0.78	0.39	–	–
	lex semi-automatic	0.27	0.78	0.40	–	–
	syn automatic	0.31	0.03	0.06	–	–
	syn semi-automatic	0.17	0.17	0.17	–	–
Test	baseline	0.11	1.0	0.19	–	–
	lex semi-automatic	0.25	0.81	0.39	0.18	3rd
	syn semi-automatic	0.18	0.81	0.29	0.19	8th

Two main experiments were carried out: the first one applies the lexical module, whereas the second one only uses syntactic information to solve implications. These runs were named *lex* and *syn*, respectively. A simple combination of both modules does not improve the results. Therefore we believe that subsequent work could be related to the combination of these modules in a collaborative way rather than by means of other simpler techniques.

Moreover, each run (*lex* or *syn*) was processed with the two types of corpus output, *automatic* and *semi-automatic*, created from the original AVE corpora. Table 1 reveals that, although the semi-automatic experiments obtain better

³ Details about the grammatical and relationship weights can be found in [5].

⁴ The proposed baseline was generated setting all pairs as VALIDATED.

results, the effort needed to generate this corpus is not worth in comparison with the gain of accuracy. The approach that achieved better results was *lex*. This is due to the fact that there are some cases where the generated hypothesis is syntactically incorrect and consequently the tree cannot be correctly generated. Future work for solving these situations will consist in refining the set of regular expressions that creates the hypotheses.

4 Conclusions and Future Work

This work contributes to establish a starting point in knowing the accuracy levels that we can obtain without semantic knowledge. However, although the results are very promising and reveal that the use of statistical measures related to string similarities and syntactic constraints are useful, we strongly believe that in many cases a fine semantic interpretation is needed in order to solve the implication correctly.

Subsequent work is oriented towards two aspects: (i) a previous preprocessing that discards inconsistent answers. We have found many cases where the calculated answer is inconsistent with the question (e.g. *a_id* number *17_3* in the development corpus, the question asks for a date and the given answer is a person name, also *a_id* number *1_5* in the test corpus is an adjective whereas its associated question requires a quantity). These situations generate a malformed hypothesis causing a wrong learning. Moreover, this preprocessing task should not be difficult since many QA systems currently include it; and (ii) the integration of semantic relations extracted from the snippets by means of resources intended for this task (e.g. knowing the entity role in a sentence would help to make semantic constraints within the entailment process; and lexical-semantic databases, such as WordNet, could establish entailment relations between verbs, detect the nouns that could accompany an adjective, is-a relations, etc).

References

1. Peñas, A., Rodrigo, A., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2007)
2. Ferrández, O., Micol, D., Muñoz, R., Palomar, M.: A Perspective-Based Approach for Solving Textual Entailment Recognition. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, June 2007, pp. 66–71. Association for Computational Linguistics, Prague (2007)
3. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out: Proceedings of the Association for Computational Linguistics Workshop, Barcelona, Spain, July 2004, pp. 74–81 (2004)
4. Lin, D.: Dependency-based Evaluation of MINIPAR. In: Workshop on the Evaluation of Parsing Systems, Granada, Spain (1998)
5. Ferrández, O., Micol, D., Muñoz, R., Palomar, M.: The Contribution of the University of Alicante to AVE 2007. In: Working Notes of the CLEF 2007 Workshop, Budapest, Hungary (September 2007)

Combining Lexical Information with Machine Learning for Answer Validation at QA@CLEF 2007

M.Á. García-Cumbreras, J.M. Perea-Ortega, F. Martínez-Santiago,
and L. Alfonso Ureña-López

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

Abstract. This document contains the description of the experiments carried out by the SINAI group. We have developed an approach based on several lexical measures integrated by means of different machine learning models. Based on lexical features it obtains a 41% of accuracy in answer validation for the Question-Answering task.

1 Introduction

This document contains the description of the experiments carried out by the SINAI group¹ at the AVE subtask of QA@CLEF 2007 [1], using English as target language. We have developed an approach based on several lexical measures integrated by means of different machine learning models. More precisely, we have evaluated three features based on lexical similarity. In order to calculate the semantic distance between a pair of tokens (stems), we have tried several measures based on Lin's similarity measure [2]. In spite of the relatively straightforward approach we have obtained a remarkable accuracy.

2 Approach Description

We have developed a system based on Machine Learning (ML) methods, which makes use of a binary classifier to solve the answer validation. We can distinguish between two processes: training and classification.

The data was given by triples (question, exact answer and supporting text passage). We used the question and the exact answer in our experiments.

In the training process we have extracted several features for each training collection². Previous results have been evaluated using the existing entailment judgements of these collections, and Machine Learning parameters have been adjusted.

¹ <http://sinai.ujaen.es>

² Answer Validation Exercise training collection and Third Recognizing Textual Entailment Challenge (RTE3) training and set collections.

We have trained the classifier obtaining a *learned model* which will be used later in the classification process.

In the classification process we extract the same features used in the training process for each pair question-answer. The classification algorithm uses these features and the *learned model* obtained in the training process and returns a boolean value (*correct* or *incorrect*) for each pair question-answer. Figure 1 describes the system architecture.

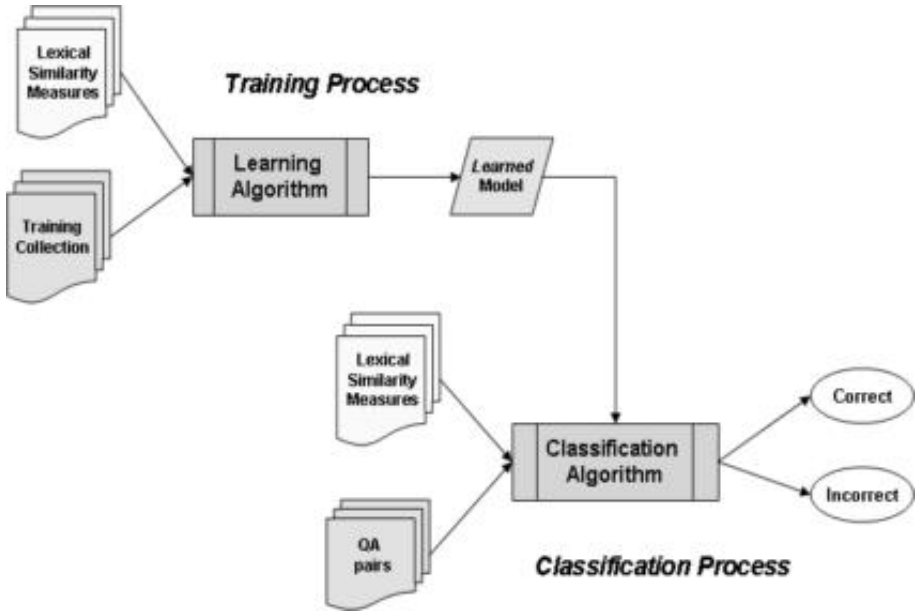


Fig. 1. System architecture

The extracted features are related to the lexical similarity. In our experiments we have applied four different lexical similarity measures, which are explained below.

2.1 Lexical Similarity

This experiment approaches the answer validation task, based on the extraction of a set of lexical measures, that check the existing similarity between the hypothesis-text pairs. Our approach is similar to [3] but the matching between pairs of words is relaxed by using the Lin's similarity measure [2] through Wordnet hierarchy. More concisely, we have applied simple matching, Binary Matching and Consecutive Subsequence Matching. In this task we have considered the answers as hypotheses and questions as texts.

Before the calculation of the different measures, the first step was to preprocess the pairs using the English *stopwords* list and the Porter stemmer available in

GATE³. In this step we also obtain the Part Of Speech (POS) of each token using GATE.

After that, we have applied four different measures or techniques:

- **Simple Matching:** this technique calculates the semantic distance between the stems of each question and its answer. If the distance exceeds a threshold, both stems are considered similar and the similarity weight value increases in one. The accumulated weight is normalized dividing it by the number of terms of the answer (hypothesis). In this experiment we have considered the threshold 0.5. The values of semantic distance measure range from 0 to 1. In order to calculate the semantic distance between two stems, we have tried several measures based on WordNet [4]. **Lin’s similarity measure** [2] was shown to be best overall measures. It uses the notion of information content and the same elements as Jiang and Conrath’s approach [5] but in a different fashion:

$$sim_L(c_1, c_2) = \frac{2 \times \log p(lso(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

where c_1 and c_2 are *synsets*, $lso(c_1, c_2)$ is the information content of their lowest super-ordinate (most specific common subsumer) and $p(c)$ is the probability of encountering an instance of a *synset* c in a specific corpus like the Brown Corpus of American English[6].

The Simple Matching technique is defined in the following equation:

$$SIM_{matching} = \frac{\sum_{i \in H} similarity(i)}{|H|}$$

where H is the set that contains the elements of the answer (hypothesis) and $similarity(i)$ is defined like:

$$similarity(i) = \begin{cases} 1 & \text{if } \exists j \in T \text{ } sim_L(i, j) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- **Binary Matching:** this measure is the same that the previous one but modifying the *similarity* function:

$$similarity(i) = \begin{cases} 1 & \text{if } \exists j \in T \text{ } i = j \\ 0 & \text{otherwise} \end{cases}$$

- **Consecutive Subsequence Matching:** this technique relies on forming subsequences of consecutive stems in the answer (hypothesis) and matching them in the question (text). The minimal size of the consecutive subsequences is two, and the maximum is the maximum size of the answer. Every correct matching increases in one the final weight. The sum of the obtained weights of the matching between subsequences of a certain size or length is normalized by the number of sets of consecutive subsequences of the answer created for this length. These weights are accumulated and normalized

³ <http://gate.ac.uk/>

by the size of the answer less one. The Consecutive Subsequence Matching technique is defined in the following equations:

$$CSS_{matching} = \frac{\sum_{i=2}^{|H|} f(SH_i)}{|H| - 1}$$

where SH_i is the set that contains the subsequences of the answer with i size or length and $f(SH_i)$ is defined like:

$$f(SH_i) = \frac{\sum_{j \in SH_i} matching(j)}{|H| - i + 1}$$

where

$$matching(i) = \begin{cases} 1 & \text{if } \exists k \in ST_i \ k = j \\ 0 & \text{otherwise} \end{cases}$$

where ST_i represents the set that contains the subsequences with i size from question (text).

- **Trigrams:** this technique relies on forming trigrams of words in the answer and matching them in the question. If a answer trigram matches in question, then the similarity weight value increases in one. The accumulated weight is normalized dividing it by the number of trigrams of the answer.

In order to obtain the results of our experiments we have used two CPAN⁴ Perl modules: the *Wordnet::Similarity* and the *Wordnet::QueryData*. We have employed the *Wordnet::QueryData* Perl module for getting the *synsets* of each (*stem, POS*) pair from the text and the hypothesis. Then, we have used the *Wordnet::Similarity* module for computing the semantic relatedness of two word senses, using the information content based measure described by Lin[2].

3 Experiments and Results

The algorithms used in the experiments as binary classifiers are two, namely, Bayesian Logistic Regression (BBR)[7] and TiMBL [8]. Both algorithms have been trained with the development data provided by the organization of the Pascal challenge (RTE-3) and the AVE task of CLEF.

As it has been explained in previous sections, a model is generated via the supervised learning process. This model is used by the classification algorithm, which will decide whether an answer is entailed by the given snippet or not.

Table 1 shows two official results and two non official, where:

- **Exp1** uses three lexical similarities (*SIMmatching* + *CSSmatching* + *Trigrams*). The model has been trained using the development data provided by the organization of the Pascal challenge, RTE-3. The ML method used was BBR.

⁴ <http://www.cpan.org>

- **Exp2** uses the same three features. The model has been trained using the development data provided by the organization of the Answer Validation Exercise task, AVE-2007, and the development data provided by the organization of the Pascal challenge, RTE-3. The ML method used was TiMBL.
- **Exp3** (non-official) uses the same three features. The model has been trained using the development data provided by the organization of the Answer Validation Exercise task, AVE-2007, and the development data provided by the organization of the Pascal challenge, RTE-3. The ML method used was BBR.
- **Exp4** (non-official) uses the same three features. The model has been trained using the development data provided by the organization of the Pascal challenge, RTE-3. The ML method used was TiMBL.

Table 1. Results with TiMBL and BBR classifiers

Experiment	Classifier	Train Data	F measure	Qa accuracy
Exp1	BBR	RTE-3	0.19	0.08
Exp2	TiMBL	RTE-3 and AVE-2007	0.37	0.41
Exp3 (non-official)	BBR	RTE-3 and AVE-2007	0.17	0.08
Exp4 (non-official)	TiMBL	RTE-3	0.25	0.32

As we expected, the best result is obtained by means of the use of both development collections, RTE-3 and AVE-2007, and the ML method TiMBL. TiMBL has been used in some classification experiments, obtaining better results than BBR [9].

4 Conclusions and Future Work

In spite of the simplicity of the approach, we have obtained remarkable results: each set of features has reported relevant information, concerning the entailment judgement determination. Our experiments approach the textual entailment task being based on the extraction of a set of lexical measures which show the existing similarity between the hypothesis-text pairs.

We have applied Simple Matching, Binary Matching, Consecutive Subsequence Matching and Trigrams, but the matching between pairs of words is relaxed by using the Lin's similarity measure through Wordnet hierarchy.

Finally, we want to implement a hierarchical architecture based on constraint satisfaction networks. The constraints will be given by the set of available features and the maintenance of the integrity according to the semantic of the phrase.

Acknowledgments

This work has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id.514 granted by the University of Jaén.

References

1. Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2007)
2. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning (1998)
3. Ferrandez, O., Mícolo, D., Muñoz, R., Palomar, M.: Técnicas léxico-sintácticas para reconocimiento de implicación textual. *Tecnologías de la Información Multilingüe y Multimodal* (in press, 2007)
4. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures (2001)
5. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics (1997)
6. Resnik, P.: Using information content to evaluate semantic similarity. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (1995)
7. Genkin, A., Lewis, D.D., Madigan, D.: BBR: Bayesian logistic regression software. Center for Discrete Mathematics and Theoretical Computer Science, Rutgers University (2005), <http://www.stat.rutgers.edu/~madigan/bbr/>
8. Daelemans, W., Zavrel, J., van der Sloot, K., van den Bosch, A.: TiMBL: Tilburg Memory Based Learner, version 1.0, Reference Guide. ILK Technical Report (1998)
9. García Cumbreras, M.A., Ureña López, A., Martínez Santiago, F.: BRUJA: Question Classification for Spanish. Using Machine Translation and an English Classifier. In: Proceedings of the MLQA 2006 (2006)

Using Recognizing Textual Entailment as a Core Engine for Answer Validation

Rui Wang¹ and Günter Neumann^{2,*}

¹ Saarland University, Saarbrücken, Germany
rwang@coli.uni-sb.de

² LT lab, DFKI, Saarbrücken, Germany
neumann@dfki.de

Abstract. This paper is about our approach to answer validation, which centered by a *Recognizing Textual Entailment* (RTE) core engine. We first combined the question and the answer into *Hypothesis* (**H**) and view the document as *Text* (**T**); then, we used our RTE system to check whether the entailment relation holds between them. Our system was evaluated on the *Answer Validation Exercise* (AVE) task and achieved f-measures of 0.46 and 0.55 for two submission runs, which both outperformed others' results for the English language.

1 Introduction and Related Work

Question Answering (QA) is an important task in *Natural Language Processing* (NLP), which aims to mine answers to natural language questions from large corpora. *Answer validation* (AV) is to evaluate the answers obtained by the former stages of a QA system and select the most proper answers for the final output.

In recent years, a new trend is to use RTE (Dagan et al., 2006) to do answer validation, cf. the AVE 2006 Working Notes (Peñas et al., 2006). Most of the groups use lexical or syntactic overlapping as features for machine learning; other groups derive the logic or semantic representations of natural language texts and perform proving.

We also developed our own RTE system, which proposed a new sentence representation extracted from the dependency structure, and utilized the Subsequence Kernel method (Bunescu and Mooney, 2006) to perform machine learning. We have achieved good results on both the RTE-2 data set (Wang and Neumann, 2007a) and the RTE-3 data set (Wang and Neumann, 2007b), especially on *Information Extraction* (IE) and QA pairs.

Therefore, the work we have done has two motivations: 1) to improve answer validation by using RTE techniques; and 2) to further test our RTE system in concrete NLP applications. The following of the paper will start with introducing our AVE system, which consists of the preprocessing part, the RTE core engine, and the post-processing part. Then, the results of our two submission runs will be shown, followed by a discussion on error sources. In the end, we will summarize our work.

* The work presented here was partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project Hy-LaP (FKZ: 01 IW F02) and by the EU-funded project QALL-ME (FP6 IST-033860).

2 Our RTE-Based AVE System

Our AVE system uses an RTE system (**TERA** – Textual Entailment Recognition for Application) as a core engine. The preprocessing module mainly adapts questions, their corresponding answers, and supporting documents into *Text(T)-Hypothesis(H)* pairs, assisted by manually designed patterns. The post-processing module will validate each answer and select a most proper one based on the output of the RTE system.

2.1 Preprocessing

The given input of the AVE task is a list of questions, their corresponding answers and the documents containing these answers. Usually, we need to validate several answers for each question. For instance, for the question, “*In which country was Edouard Balladur born?*” the QA system gives out several candidate answers to this question, “*Frances*”, “*12% jobless rate*”, or “*7*”, and also supporting documents where the answers come from. Here, the assumption for the validation process is that *if the answer is to the question, the document which contains the answer should entail the statement derived by combining the question and the answer.*

In order to combine a question and an answer into a statement, we manually constructed some language patterns for the input questions. As for the question given above, we will apply the following pattern, “*Edouard Balladur was born in <Answer>*”, and substitute the “*<Answer>*” by each candidate answer to form an **H** – a hypothesis. Since the supporting documents are naturally the **Ts** – texts, the **T-H** pairs are built up accordingly, and these **T-H** pairs can be the input for any generic RTE systems.

2.2 The RTE Core Engine

The RTE core engine contains a main approach with two backup strategies (Wang and Neumann, 2007a). In brief, the main approach firstly extracts common nouns between **T** and **H**; then it locates them in the dependency parse tree as *Foot Nodes* (FNs). Starting from the FNs, a common parent node can be found in each tree, which will be named as *Root Node* (RN); Altogether, FNs, the RN, and the dependency paths in-between will form a *Tree Skeleton* (TS) for each tree. On top of this feature space, we can apply subsequence kernels to represent these TSs and perform kernel-based machine learning to predict the final answers discriminatively.

The backup strategies will deal with the **T-H** pairs which cannot be solved by the main approach. One backup strategy is called Triple Matcher, as it calculates the overlapping ratio on top of the dependency structures in a triple representation¹; the other is simply a Bag-of-Words (BoW) method, which calculates the overlapping ratio of words in **T** and **H**.

2.3 Post-processing

The RTE core engine has given us: 1) for some of the **T-H** pairs, we directly know whether the entailment holds; 2) every **T-H** pair has a triple similarity score and a

¹ A triple is of the form <node1, relation, node2>, where node1 represents the head, node2 the modifier, and relation the dependency relation.

BoW similarity score. If the **T-H** pairs are covered by our main approach, we will directly use the answers; if not, we will use a threshold to decide the answer based on the two similarity scores. In practice, the thresholds are learned from the training corpus.

For adapting the results back to the AVE task, the “YES” entailment cases will be the validated answers and the “NO” entailment cases will be the rejected one. In addition, the selected answers (i.e. the best answers) will naturally be the pairs covered by our main approach or (if not,) with the highest similarity scores.

3 Results and Error Analysis

We have submitted two runs for AVE2007. Both of the two runs we have used the main approach plus one backup strategy. In the first run, the BoW similarity score was the backup, while in the second run, the triple similarity score was taken. We have used *Minipar* (Lin, 1998) as our dependency parser and our machine learning process was performed by the classifier SMO from the WEKA toolkit (Witten and Frank, 1999). The following Table 1 shows the results,

Table 1. Results of our two submission runs

Submission Runs	Recall	Precision	F-measure	QA Accuracy
dfki07-run1.txt	0.62	0.37	0.46	0.16
dfki07-run2.txt	0.71	0.44	0.55	0.21

Though the absolute scores are not very promising, they are still better than all the others’ results for the English language this year. The second run outperforms the first run in all respects, which shows advantages of the triple similarity score. The gold standard does not contain the “*SELECTED*” answers, thus, we will not discuss the QA accuracy here. Instead, the error analysis will focus on the loss of recall and precision.

As for recall, among all the errors, half of them belong to one type. For questions like “*What is the occupation of Kiri Te Kanawa?*” we have used the pattern “*The occupation of Kiri Te Kanawa is <Answer>*”, which has caused problems, because “*occupation*” usually did not appear in the documents. Instead, a pattern like “*Kiri Te Kanawa is <Answer>*” might be much better. Some other errors are from the noise of web documents, on which the dependency parser could not work very well.

The precision of our two runs are rather poor. After taking a closer look at the errors, we have found that most of the errors also belong to one type. In those answer-document pairs (e.g. id=119_2, id=125_1, id=133_1, etc.), the answers are usually very long, which consist of a large part of the documents. Some extreme cases (e.g. id=112_2, id=172_2, etc.), the answers are very long and exactly the same as the documents. Due to the characteristics of our method (i.e. using RTE for AVE), these answers will get high similarity scores, which are wrongly validated.

In addition, some other errors like trivial answers (e.g. “*one*”) could be avoided by adding some rules. As a whole, more fine-grained classification of answers could be helpful to improve the system.

4 Conclusion and Future Work

In conclusion, we have described our approach to answer validation using RTE as a core engine. On the one hand, it is an effective way to do the answer validation task; on the other hand, it is also a promising application for our developed RTE system. The results have shown the advantages of our combination of the main approach and backup strategies.

After error analysis, the possible future directions are: 1) preprocessing the documents to clean the noisy web data; 2) making the patterns be automatically generated; 3) utilizing question analysis tools to acquire more useful information.

References

1. Bunescu, R., Mooney, R.: Subsequence Kernels for Relation Extraction. In: *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge (2006)
2. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) *MLCW 2005. LNCS (LNAI)*, vol. 3944, pp. 177–190. Springer, Heidelberg (2006)
3. Lin, D.: Dependency-based Evaluation of MINIPAR. In: *Workshop on the Evaluation of Parsing Systems* (1998)
4. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the Answer Validation Exercise 2006. In: *AVE 2006 Working Notes* (2006)
5. Wang, R., Neumann, G.: Recognizing Textual Entailment Using a Subsequence Kernel Method. In: *Proc. of AAAI 2007* (2007a)
6. Wang, R., Neumann, G.: Recognizing Textual Entailment Using Sentence Similarity based on Dependency Tree Skeletons. In: *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007, pp. 36–41 (2007b)
7. Witten, I.H., Frank, E.: *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco (1999)

A Supervised Learning Approach to Spanish Answer Validation

Alberto Téllez-Valero, Manuel Montes-y-Gómez,
and Luis Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica
Grupo de Tecnologías del Lenguaje

Luis Enrique Erro no. 1, Sta. María Tonantzintla, Pue.; 72840; Mexico
{albertotellezv,mmontesg,villasen}@inaoep.mx

Abstract. This paper describes the results of the INAOE’s answer validation system evaluated at the Spanish track of the AVE 2007. The system is based on a supervised learning approach that considers two kinds of attributes. On the one hand, some attributes indicating the textual entailment between the given support text and the hypothesis constructed from the question and answer. On the other hand, some new features denoting certain answer restrictions as imposed by the question’s type and format. The evaluation results were encouraging; they reached a F-measure of 53% (the best performance in the Spanish track), and outperformed the standard baseline by 15 percentage points.

1 Introduction

Given a question, a candidate answer and a support text, an answer validation system must decide whether accept or reject the candidate answer. In other words, it must determine if the specified answer is correct and supported.

In the previous answer validation exercise (AVE 2006), the answer validation systems were based on the idea of recognizing the textual entailment between the support text and an affirmative sentence (called hypothesis) created from the combination of the question and answer. In order to accomplish this recognition, these systems probed several approaches, ranging from simple ones taking advantage of lexical overlaps to more complexes founded on the use of a logic representation [1].

The approach based on lexical overlaps is quite simple, but surprisingly it has achieved very competitive results. Representative methods of this approach determine that H (the hypothesis) is entailed from T (the support text) only considering characteristics such as named entity overlaps [2], n-gram overlaps and the size of the longest common subsequence (LCS) [3].

The simplicity is the strength of this approach but at the same time is its weakness. All overlap-based methods have problems to deal with situations where the answer should be satisfy simple type restrictions imposed by the question. For instance, the candidate answer “*Javier Sotomayor*” is clearly incorrect for the question “*What is the world record in the high jump?*”, but it will be validated

as accepted because the high lexical similarity between the formed hypothesis “*The world record in the high jump is Javier Sotomayor*” and the corresponding support text “*The world record in the high jump, obtained by Javier Sotomayor, is 2.45 meters.*”.

The proposed system adopts several ideas from recent systems (in particular from [2,3]): it is based on a supervised learning approach that considers a combination of some previously-used features. However, in addition, it also includes some new characteristics that allow tackling the discussed problem.

2 The Answer Validation System

In resume, the main characteristics of our system are the following:

1. It only considers content words for computing word overlaps and LCS.
2. It uses POS tags for the calculus of the LCS.
3. It makes a syntactic transformation of the generated hypothesis in order to simulate the active and passive voices.
4. It applies some manually-constructed lexical patterns to help treating support texts containing an apposition and adjectival phrases.
5. It includes some new features denoting certain answer restrictions as imposed by the question’s class.

For a complete description of the system refer to [4].

3 Experimental Evaluation

3.1 Training and Test Sets

In order to avoid the low recall in the validated answers we assembled a more balanced training set. Basically, we joined some answers from the training sets of the AVE 2006 and 2007. This new training set contains 2022 answers, where 44% are validated and 56% rejected. On the other hand, the evaluation set for the Spanish AVE 2007 contains 564 answers (22.5% validated and 77.5% rejected) corresponding to 170 different questions.

3.2 Results

This year we submitted two different runs considering two different classification algorithms. The first run (RUN 1) used a single support vector machine classifier, whereas the second run (RUN 2) employed an ensemble of this classifier based on the AdaBoostM1 algorithm.

Table 1 shows the evaluation results corresponding to our two submitted runs. It also shows (in the last row) the results for a 100% VALIDATED baseline (i.e., an answer validation system that accepted all given answers). The results indicate that our methods achieved a high recall and a middle level precision, which means that they correctly accepts must of the right answers (there are

a few false negatives), but also incorrectly accepts some wrong ones (there are several false positives).

An analysis of false positives shows us that the main problem of our approach is still the high overlap that exists between the T and H although the evaluated answer is wrong. For instance, in the question “*Who made Windows 95?*”, the wrong candidate answer “*business*” is validated as accepted. This error occurs because the content terms in the formed hypothesis “*business made Windows 95*” can be totally overlap by the support text “*Windows 95 is the new version of the operating system made for the business Microsoft, . . .*”. These cases evidenced the necessity of including more information into the overlap checking process, such as term dependencies and more restrictive data about the kind of expected answer.

Table 1. General evaluation of the INAOE’s system (here TP, FP, TN, and FN refers to true positives, false positives, true negatives, and false negatives, respectively)

	TP	FP	TN	FN	Precision	Recall	F-measure
RUN 1	109	176	248	18	0.38	0.86	0.53
RUN 2	91	131	293	36	0.41	0.72	0.52
100% VALIDATED	127	424	–	–	0.23	1	0.37

This year the AVE organizers decide to include a new evaluation measure, called qa-accuracy. This measure allows evaluating the influence of the answer validation systems into the question answering task. In order to compute this measure the answer validation systems must select only one validated answer for each question. This way, the qa-accuracy expresses the rate of correct selected answers. Table 2 presents the qa-accuracy results of our two runs. It also shows (in the last row) the best results obtained at QA@CLEF 2007 for the same set of questions.

Table 2. Evaluation results obtained by the qa-accuracy measure

	Total	Selected answers			QA-accuracy
		Right	Wrong	Inexact	
RUN 1	129	76	47	6	0.45
RUN 2	107	62	40	5	0.36
BEST QA SYSTEM	–	84	–	–	0.49

In order to do a detail evaluation of our system we also measured its precision over the subset of 101 questions that have at least one correct candidate answer. In this case, RUN 1 selected the right candidate answer for 75% of the questions, and RUN 2 for 61%. For the rest of the questions (69 questions), for which no correct candidate answer exists, RUN 1 correctly answered NIL in 49% of the cases, whereas the RUN 2 correctly responded NIL in 61% of the questions.

It is important to mention that current qa-accuracy measure does not take into account the correctly selected NIL answers. That is, it does not consider NIL answers as correct answers for any question (even for those cases that do not have the answer in the test document collection). Considering NIL answers into the evaluation, our answer validation system – in the RUN 1 – could reach an accuracy equal to the best QA system (i.e., 49%).

4 Conclusions

This paper presents the evaluation results of the INAOE's answer validation system at the Spanish track of the AVE 2007. Our system adopts several ideas from recent overlap-based methods; basically, it is based on a supervised learning approach that uses a combination of some previous used features, in particular, word overlaps and longest common subsequences. However, it also includes some new notions that extend and improve these previous methods.

The evaluation results are encouraging; they show that the proposed system achieved a 53% of F-measure, obtaining the best result in the Spanish track. As future work we plan to enhance the question-answer compatibility analysis as well as to apply other attributes in the supervised learning process.

Acknowledgments

This work was done under partial support of CONACYT (project grant 43990 and scholarship 171610). We also thank the CLEF organizers.

References

1. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the answer validation exercise 2006. In: [5], pp. 257–264
2. Rodrigo, Á., Peñas, A., Herrera, J., Verdejo, F.: The effect of entity recognition on answer validation. In: [5], pp. 483–489
3. Kozareva, Z., Vázquez, S., Montoyo, A.: University of Alicante at QA@CLEF: Answer validation exercise, In: [5], pp. 522–525
4. Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L.: INAOE at AVE 2007: Experiments in Spanish answer validation. In: Working notes for the 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19–21 (2007)
5. Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.): CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)

UAIC Participation at AVE 2007

Adrian Iftene¹ and Alexandra Balahur-Dobrescu^{1,2}

¹ UAIC: Faculty of Computer Science, “Alexandru Ioan Cuza” University, Romania

² University of Alicante, Department of Software and Computing Systems,

Alicante, Spain

{adiftene, abalahur}@info.uaic.ro

Abstract. Textual Entailment Recognition (RTE) is a recently proposed task, aiming at capturing the means through which textual inferences can be made. Moreover, using such a module is meant to contribute to the increase in performance of many NLP applications, such as Summarization, Information Retrieval or Question Answering, both for answer ranking as well as for answer validation. This article presents the manner in which we used the TE system built for the RTE3 competition this year for the AVE exercise. We describe the steps followed in building the patterns for question transformation, the generation of the corresponding hypotheses and finally for answer ranking. We conclude by presenting an overview of the performance obtained by this approach and a critical analysis of the errors obtained.

1 Introduction

AVE¹ is a new task introduced in the QA@CLEF competition, with the aim of promoting the development and evaluation of subsystems validating the correctness of the answers given by QA systems. In this task, the systems must emulate human assessment of QA responses and decide whether an Answer to a Question is correct or not according to a given Text.

Participant systems receive a set of triplets (Question, Answer, and Supporting Text) and they must return a Boolean value for each triplet. Results are evaluated against by the QA human assessments.

The AVE task described is hence similar to the Recognizing Textual Entailment task (Dagan et al., 2006), for which systems are built and a Recognizing Textual Entailment competition has been organized for the last three years by the PASCAL excellence network for multimodal interfaces.

Formally, textual entailment is defined (Dagan et al., 2006) as a directional relation between two text fragments, termed T - the entailing text, and H - the entailed text. It is then said that T entails H if, typically, a human reading T would infer that H is most likely true. Textual entailment recognition is thus the task of deciding, given T and H , whether T entails H . In the textual entailment competition, participants are provided

¹ <http://nlp.uned.es/QA/ave/>

with pairs of small text snippets and they must build a system that should judge the truth value of the entailment relation for each pair.

2 Textual Entailment System

The main idea of our system is to transform the hypothesis making use of extensive semantic knowledge from sources like DIRT, WordNet, Wikipedia, and database of acronyms. Additionally, we built a system to acquire the extra Background Knowledge needed and applied complex grammar rules for rephrasing in English.

2.1 Tools

LingPipe

The first step splits the initial file into pairs of files of text and hypothesis. All these files are then sent to the LingPipe² module in order to find the Named entities.

MINIPAR

In parallel, we transform with MINIPAR³ (Lin, 1998) both the text and the hypothesis into dependency trees. For every node from the MINIPAR output (which represents a simple word belonging to a sentence), we consider a stamp called **entity** with three main features: the node lemma, the father lemma and the edge label.

2.2 Resources

The main module receives separate files for each text and hypothesis pair from the initial data (test or development). The remaining resources used are DIRT, WordNet, the Acronyms Database and the Background Knowledge. They are used to expand each remaining word from the hypothesis to a list of similar or related terms and thus increase the probability to find the initial word or any of its equivalents in the list of words from the text.

The DIRT resource: DIRT⁴ (Discovery of Inference Rules from Text) is both an algorithm and a resulting knowledge collection created by Lin and Pantel at the University of Alberta (Lin and Pantel, 2001), (Lin, 2001). A path, extracted from a parse tree, is an expression that represents a binary relationship between two nouns. For the hypothesis verbs in the MINIPAR output without correspondence, we extract templates with DIRT like format. In the same way, we build a list with templates for the verbs in the text tree. With these two lists of templates we perform a search in the DIRT database and extract the “best” trimming using template type (full or partial) and the DIRT score.

² <http://www.alias-i.com/lingpipe/>

³ <http://www.cs.ualberta.ca/~lindek/minipar.htm>

⁴ http://aclweb.org/aclwiki/index.php?title=DIRT_Paraphrase_Collection

eXtended WordNet: eXtended WordNet⁵ is an ongoing project at the Human Language Technology Research Institute, University of Texas at Dallas. In the eXtended WordNet, the WordNet glosses are syntactically parsed transformed into logic forms and content words are semantically disambiguated. For non-verbs nodes from the hypothesis tree, if in the text tree we do not have nodes with the same lemma, we search for their synonyms in the extended WordNet.

The **acronyms' database**⁶ helps our program to find relations between the acronym and its meaning: "US - United States".

The **Background Knowledge** was used in order to expand the named entities and the numbers. It was built semi-automatically, and it used a module in which language could be set according to the current system working language and thus the corresponding Wikipedia⁷ could be selected. For every named entity or number in the hypothesis, the module extracted from Wikipedia a file with snippets with information related to them.

For every node transformed with DIRT or with the eXtended WordNet, we consider its **local fitness** as being the similarity value indicated by DIRT or by eXtended WordNet. In other cases, when there is a direct mapping or when we use the acronyms database or the Background Knowledge, we consider the **local fitness** of the node to be 1.

2.3 Rules

Semantic Variability Rules: negations and context terms

For every verb from the text and hypothesis we consider a Boolean value which indicates whether the verb has a negation or not, or, equivalently, if it is related to a verb or adverb **diminishing** its sense or not. For that, we use the POS-tags and a list of words we consider as introducing a negation: "no", "don't", "not", "never", "may", "might", "cannot", "should", "could", etc. For each of these words we successively negate the initial truth-value of the verb, which by default is "false". The final value depends on the number of such words.

Rule for Named Entities from hypothesis without correspondence in text Additionally, we have a separate rule for named entities from the hypothesis without correspondence in the text. If the word is marked as named entity by LingPipe, we try to use the acronyms' database or obtain information related to it from the background knowledge. In the event that even after these operations we cannot map the word from the hypothesis to one word from the text, we increase a value that counts the problems regarding the named entities in the current pair. We then proceed to calculating a fitness score measuring the syntactic similarity between the hypothesis and the text, further used as one of the features that the two classifiers used are trained on.

⁵ <http://xwn.hlt.utdallas.edu/>

⁶ <http://www.acronym-guide.com>

⁷ http://en.wikipedia.org/wiki/Main_Page

Rule for determination of entailment

After transforming the hypothesis tree, we calculate a global fitness score using the following **extended local fitness** value for every node from the hypothesis - which is calculated as sum of the following values:

- 1. local fitness obtained after the tree transformation and node mapping,
- 2. parent fitness after parent mapping,
- 3. mapping of the node edge label from the hypothesis tree onto the text tree,
- 4. node position (left, right) towards its father in the hypothesis and position of the mapping nodes from the text.

We calculate for every node from the hypothesis tree the value of the extended local fitness, and afterwards consider the normalized value relative to the number of nodes from the hypothesis tree. We denote this result by *TF* (*Total Fitness*). After calculating this value, we compute a value *NV* (the negation value) indicating the number of verbs with the same value of negation. Because the maximum value for the extended fitness is 4, the complementary value of the *TF* is $4 - TF$. The formula for the **global fitness** used is therefore:

$$GlobalFitness = NV * TF + (1 - NV) * (4 - TF)$$

Using the development data, we establish a threshold value of 2.06, and according to this, we decide that pairs above it will have the answer “yes” for entailment.

2.4 Results in RTE3

We submitted two runs for our system, with the difference residing in the parameters used in calculating the extended local fitness, but with the same final score: 69.13 %.

To be able to see each component’s relevance, the system was run in turn with each component removed. The results in the table below show that the system part verifying the NEs is the most important.

Table 1. Components relevance

System Description	Precision	Relevance
Without DIRT	0.6876	0.54 %
Without WordNet	0.6800	1.63 %
Without Acronyms	0.6838	1.08 %
Without BK	0.6775	2.00 %
Without Negations	0.6763	2.17 %
Without NEs	0.5758	16.71 %

Twenty-six teams participated in the third challenge, and even though this was our first participation in the RTE competition, our system was ranked third, being among the best in the competition.

3 Using the TE System in the AVE track

The data provided in the AVE task are in the following format:

Table 2. Question 148 from the test data

<q id="148" lang="EN"> <q_str>When was Yitzhak Rabin born?</q_str> <a_str>March 1 1922</a_str> <t_str doc="GH951109-000097"> Yitzhak Rabin, Prime Minister of Israel; born Jerusalem, March 1, 1922, died Tel Aviv, November 4, 1995. ... </t_str> <a_str>1992-1995</a_str> <t_str doc="en/p03/368881.xml">Yitzhak Rabin 1992-1995</t_str> <a_str>4 19</a_str> <t_str doc="168208.xml">Yitzhak Shamir Yitzhak Shamir 7 יצחק שמיר Prime Minister of Israel </t_str> <a_str>1995</a_str> <t_str doc="650871.xml">Ministry of Interior (Israel)</t_str> </q>

The system architecture is presented below:

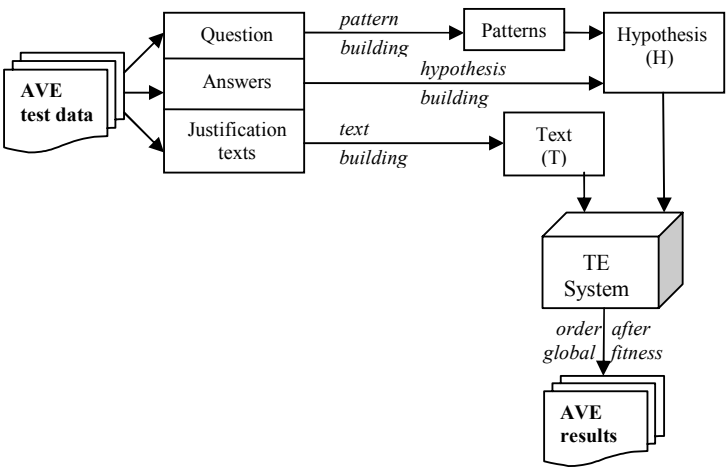


Fig. 1. AVE System

The steps executed by our system are as follows:

- We build a pattern with variables for every question according to the question type;
- Using a pattern and all possible answers, we build a set of hypotheses for each of the questions: H_1 , H_2 , H_3 etc.;
- We assign the justification snippet the role of text T and we run the TE system for all obtained pairs: (T_1, H_1) , (T_2, H_2) , (T_3, H_3) , etc.

Lastly, we consider the correct answer for the current question the candidate from the hypothesis for which we obtain the greatest global fitness.

3.1 Pattern Building

In order to use the TE system for ranking the possible answers in the AVE task, all these questions are first transformed according to the algorithm presented in (Bar-Haim et al., 2006).

For question 148 we have:

Question: *When was Yitzhak Rabin born?*

Our program generates the following pattern:

Pattern: *Yitzhak Rabin was born at DATE.*

where *DATE* is the variable in this case. We generate specific patterns according to the following answer types: Measure, Person, Other, Location, Organization and Date. Next table presents the identified types of patterns:

Table 3. Examples of Patterns

Answer type	Cases Number	Question example	Pattern
Date	3	When was Yitzhak Rabin born?	Yitzhak Rabin was born at DATE .
Measure	5	How many bush fires were there near Sydney in January 1994?	MEASURE bush fires were there near Sydney in January 1994.
Location	6	Where is the Hermitage Museum?	The Hermitage Museum is in LOCATION .
Person	12	Who wrote the song "Dancing Queen"?	PERSON wrote the song "Dancing Queen".
Organization	17	What company makes Ribena?	ORGANIZATION company makes Ribena.
Other	23	What is Gulf War Syndrome?	Gulf War Syndrome is OTHER .

Following the building of the pattern, we proceed to constructing the corresponding hypotheses.

3.2 Hypothesis Building

Using the pattern building mechanism above and the answers provided within the AVE data, we built the corresponding hypotheses. For example, for question 148, we built, according to the answers from the English test data (“a_str” tags), the following hypotheses:

H_{148_1} : *Yitzhak Rabin was born at March 1 1922.*

H_{148_2} : *Yitzhak Rabin was born at 1992-1995.*

H_{148_3} : *Yitzhak Rabin was born at 4 19.*

H_{148_4} : *Yitzhak Rabin was born at 1995.*

For each of these hypotheses, we consider as having the role of text T the corresponding justification text (content of “t_str” tag).

3.3 Answers Classification

We consider the pairs built above as input for our Textual Entailment system. After running the TE system, the global fitness values for these pairs are the following:

$GlobalFitness(H_{148\_1}, T_{148\_1}) = 2.1148$

$GlobalFitness(H_{148\_2}, T_{148\_2}) = 1.8846$

$GlobalFitness(H_{148\_3}, T_{148\_3}) = 2.1042$

$GlobalFitness(H_{148\_4}, T_{148\_4}) = 1.7045$

Since in the considered case the highest value is obtained for the answer *March 1 1922*, we consider it as the **SELECTED** answer and the rest as **VALIDATED**. The **REJECTED** answers were considered the pairs for which we have NE problems (in this case, the global fitness has the minimum value, i.e. 0).

3.4 Results and Errors Analysis

Our AVE system has the following results:

Table 4. AVE Results

F measure	0.34
Precision over YES pairs	0.21
Recall over YES pairs	0.81
QA accuracy	0.21

Our results as compared to other participants are presented below (Peñas et al., 2007):

We compare our results against the gold file and count to see which class of answers our correct and incorrect answers pertain to. In table below we can observe the fact that most problems arose within the **REJECTED** class. The cause of this issue was that our TE system considers as being rejected only the pairs for which NE problems were encountered (in this case, the global fitness is zero). This rule functions very well for 111 cases from 174, but as it can be observed, it is not enough. In all

Table 5. Comparing AV systems performance with QA systems in English for first 5 systems

Group	QA accuracy	% of perfect selection
DFKI 2	0.21	70%
UAIC Iasi	0.21	70%
UA 2	0.19	65%
U.Indonesia	0.18	60%
UA 1	0.18	60%

other cases, we calculate the global fitness, and the answer with the highest score is considered SELECTED and all other answers are considered as VALIDATED. One solution for this problem was to train our TE system on the AVE development data and identify a specific threshold according to the AVE input data.

Table 6. Distribution on answers classes of our Results

Answers Class in Gold file	Unknown	Valid`ated	Rejected	Total
Correct		17	111	128
Incorrect	7	4	63	74

4 Conclusions

We showed how the TE system used in the RTE3 competition can successfully be used as part of the AVE system, resulting in improved ranking between the possible answers, especially in the case of questions with answers of type Person, Location, Date and Organization.

The main problem of our system arises from the rule that identifies the REJECT cases in the AVE competition. We notice that the rule regarding the presence of NEs is very good in this case and identifies 64 % of the correct cases, but it is not enough to identify the entire class of REJECTED answers. In order to better identify these situations, we must additional rules must be added in order to bring the system improvement.

References

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 Recognizing Textual Entailment, Venice, Italy (2006)

Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela, J., et al. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 177–190. Springer, Heidelberg (2006)

- Lin, D.: Dependency-based Evaluation of MINIPAR. In: Workshop on the Evaluation of Parsing Systems, Granada, Spain (May 1998)
- Lin, D.: LaTaT: Language and Text Analysis Tools. In: Proc. Human Language Technology Conference, San Diego, California (March 2001)
- Lin, D., Pantel, P.: DIRT - Discovery of Inference Rules from Text. In: Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD 2001), San Francisco, CA, pp. 323–328 (2001)
- Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Working Notes of CLEF 2007, Budapest, Hungary, 19-21 September (2007)

UNED at Answer Validation Exercise 2007

Álvaro Rodrigo, Anselmo Peñas, and Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED
{alvarory, anselmo, felisa}@lsi.uned.es

Abstract. The objective of the Answer Validation Exercise (AVE) 2007 is to develop systems able to decide if the answer to a question is correct or not. Since it is expected that a high percentage of the answers, questions and supporting snippets contain named entities, the paper presents a method for validating answers that uses only information about named entities. The promising results encourage us to improve the system and use it as a component of other systems.

1 Introduction

The Answer Validation Exercise (AVE) 2007 [4] of the Cross Language Evaluation Forum (CLEF) 2007 aims at developing systems able to decide whether the responses of a question answering (QA) system are correct or not. As a difference with last year [5], the organization does not provide the participants with text-hypothesis pairs. This year, participant systems receive a set of answers and their corresponding supporting snippets grouped by question. Besides, this year it is not mandatory to use textual entailment. Systems must return a value **VALIDATED** or **SELECTED** if they considered that the answer is correct and the snippet supports it, and return **REJECTED** if the answer is incorrect or it is not supported by the text snippet.

The system we have presented is based on the one we used at AVE 2006 [7], which gave good results in Spanish, and the one used in our participation at RTE-3 [6], which obtained also good results in textual entailment over pairs from QA. These two systems were based on named entities (NE). However, these systems needed text-hypothesis pairs that are not given at AVE 2007. This paper shows a system based on named entities that has been adapted to the new specifications at AVE 2007. We have participated with this system in English and Spanish.

Our main motivation for using named entities is the high percentage of factoid questions in QA@CLEF (e.g. 79% of questions in last year Spanish test set [3]). The answers to these questions are expected to be named entities (person names, locations, numbers, dates...) and it is expected that these questions, answers and snippets contain a high amount of named entities.

The main components of the system are described in Section 2. The results and the error analysis are shown in Section 3. Finally, some conclusions and future work are given in Section 4.

2 System Description

The system receives a set of triplets (question, answer, snippet) and decides, using only information about named entities, if the answer to the question is correct and the text snippet supports it.

As the system uses only information about named entities, the first step is to detect them in a robust way. Then, the second step is the definition and implementation of an entailment relation between named entities.

The following subsections describe in detail the steps involved in the decision of named entity entailment.

2.1 Named Entity Recognition

Numeric expressions (NUMEX), proper nouns (PN) and time expressions (TIMEX) of questions, answers and snippets are tagged using the FreeLing [1] Name Entity Recognizer (NER). The values of numeric and time expressions are also normalized in order to simplify the entailment decision.

In order to avoid errors in the process of named entity entailment, as it is explained in [7], all named entities receive the same tag NE ignoring the named entity categorization given by the tool.

2.2 Named Entity Entailment

Once the named entities of questions, answers and snippets are detected, the next step is to determine the entailment relations between them. As it is explained in [6], we consider that a named entity NE1 entails a named entity NE2 if the text string of NE1 contains the text string of NE2. However, some characters change in different expressions of the same named 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 a modified entailment decision process taking into account the edit distance of Levenshtein [2]. Thus, if two named entities differ in less than 20%, then we assume that exists an entailment relation between them.

2.3 Validation Decision

In [6] and [7], we detected the entailment relation between named entities in the text and in the hypothesis. In AVE 2007 [4], this is not possible due to the fact that no hypothesis is given.

As it is described in [5], the hypotheses given by the AVE 2006 organization were build as a combination of questions and answers. This fact led us to think the possibility of developing a module able to build hypotheses with answers and questions as input. However, as our system needs only the named entities from the hypothesis, we studied how to obtain them without building a textual hypothesis. Our intuition was that the named entities of a certain hypothesis were the same as the named entities of the question plus the named entities of the answer from which the hypothesis is generated.

A look at AVE 2006 corpus shows us that our intuition was correct as Fig. 1 shows. In the example showed in the figure, the hypothesis has been obtained from the question and answer of the example. The named entities of the hypothesis (Iraq,

Question: Which country did <NE>Iraq</NE> invade in < NE >1990? </NE>
Answer: <NE>Kuwait</NE>
Hypothesis: <NE>Iraq</NE> invaded the country of <NE>Kuwait</NE> in
 <NE>1990</NE>

Fig. 1. An example of how the NEs of the hypothesis are the NEs of the question and the answer

Kuwait and 1990) correspond to named entities in the question (Iraq, 1990) and the answer (Kuwait).

Thus, the validation decision for each triplet (question, answer, snippet) is obtained taking into account the named entities from the text snippet in one way and the named entities from the question plus the named entities from the answer in another way as named entities of a supposed hypothesis.

Then, for taking the final decision, we think that in textual entailment all the elements in the hypothesis must be entailed by elements of the supporting text. Therefore, the system assumes that if there is a named entity in the hypothesis not entailed by one or more named entities in the text, then the answer is not supported or incorrect and then the system must return the value REJECTED for this triplet.

However, in pairs where all the entities in the hypothesis are entailed, there is not enough evidence to decide if the answer is correct or not. In this situation, in order to perform an experiment to obtain some information about the system performance, we decided to return the value VALIDATED.

Even though the validation decision described above shows a good performance in the Spanish development set, the results in English were lower mainly due to errors in the recognition of named entities in the text snippets. An example of these errors is shown in Fig. 2, where Italy has not been recognized as a named entity in the text snippet.

Question: What is the name of the national airline in <NE>Italy</NE>?
Snippet: Italy's national airline <NE>Alitalia</NE>

Fig. 2. An example of a NE recognition error

Then, we thought that in our validation decision process it was important that the named entities of the hypothesis (combination of question and answer) were entailed by elements in the text snippet, without the necessity that these elements were recognised as named entities. In order to study this approach, an experiment was performed over the English development set with two different systems:

1. A system that takes the validation decision as it has been explained above.
2. A system that returns REJECTED if no token (or consecutive tokens) of the text entails some named entity in the hypothesis taking the idea of entailment described in section 2.2.

The experiment results (see Table 1) show that the second system achieves a slight improvement in F measure, that is the measure used for comparing AVE systems [5]. Then, the second option of validation decision was taken for English triplets.

Table 1. Comparing validation decisions

	F
System 1	0.3
System 2	0.33

2.4 Selection Decision

In AVE 2007 [4], a new measure called qa_accuracy has been proposed to compare the results of AVE participants with the results of QA participants. The objective is to measure the performance of the answer validation system selecting an answer from a set of answers to the same question. For this purpose, it is mandatory in the task that when a system returns the value VALIDATED for one or more answers to the same question, one of them has to be tagged as SELECTED.

The system we have presented does not have a way to decide which one of the answers given as VALIDATED is the most probable to be correct. Then, we did not have an objective method for selecting answers to compare our system with the QA participants. For this reason, we decided to use a non-informative method that tagged as SELECTED the first answer of each question that is detected as correct for our system.

3 Results and Error Analysis

The described system has been tested on the Spanish and English test sets of AVE 2007. Table 2 and Table 3 show the precision, recall and F measure over correct answers obtained in English and Spanish, respectively, with a baseline system that returns VALIDATED for all answers.

In both languages, the results obtained have been better than the baselines, achieving a high recall.

Table 2. Results in English

	F	precision	recall
UNED system	0.34	0.22	0.71
100% VALIDATED baseline	0.19	0.11	1

Table 3. Results in Spanish

	F	precision	recall
UNED system	0.47	0.33	0.82
100% VALIDATED baseline	0.37	0.23	1

The errors detected in triplets where the system returns **VALIDATED** were due to the lack of knowledge. In these pairs, all the named entities from the question and the answer are entailed for some named entities in the text snippet. However, the answer is incorrect as for example the answer in Fig. 3 where the expected answer is an instrument, but the given answer is a year. Then, if the named entities of the question and the answer are entailed, our system would return **VALIDATED**.

Question: What instrument did Swann play in the duo Flanders and Swann?

Answer: 1964

Fig. 3. Example of a false **VALIDATED** answer

In some of the errors in triplets where the system returns **REJECTED**, a full name of a person (for example Steve Fosset) appeared in the question and the answer was judged as correct, but in the snippet appeared only the last name of this person (Fosset in the previous example). Our system cannot find a named entity in the text snippet that entails the full name and hence it returns **REJECTED**. As it is not certain that the person in the text was the same as in the question, we think that maybe this kind of answers should be assessed as unsupported (and then in AVE as **REJECTED**).

Regarding the measure of *qa_accuracy*, Table 4 and Table 5 show the results obtained in English and Spanish, respectively, compared with the value obtained in a perfect selection and a baseline system that validates 100% of the answers and selects randomly one of them. In the last column, the normalization of *qa_accuracy* by the perfect selection value is given.

Table 4. *qa_accuracy* results in English

	qa_accuracy	Normalized
Perfect selection	0.3	100%
UNED	0.16	55%
Random selection	0.1	35%

Table 5. *qa_accuracy* results in Spanish

	qa_accuracy	Normalized
Perfect selection	0.59	100%
UNED	0.42	70.3%
Random selection	0.25	41.45%

Even though the system uses a non-informative method for selecting answers, the results are between a perfect and a random selection.

4 Conclusions and Future Work

We have presented a system based on textual entailment that does not need to build textual hypotheses. The system uses only information about named entities and obtains results very promising. These results encourage us to use information about named entities in more complex answer validation systems.

We consider that the information about named entities can be used in two different ways:

1. As additional information in another answer validation system.
2. As a filter before using another answer validation system. Our system would reject answers that considers as incorrect and another system would take the decision in the rest of the answers. This idea arise from the fact that our system is focused on detecting incorrect answers achieving a precision of 95% and 90% for REJECTED answers in English and Spanish, respectively.

Future work is focused in improving the named entity recognition and the decision of entailment. In this way, next step is to be able of detecting the equivalence between some named entities and their acronym (for example, UN is equivalent to United Nations).

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Technology within the Text-Mess-INES project (TIN2006-15265-C06-02), the Education Council of the Regional Government of Madrid and the European Social Fund.

References

1. 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 2004), Lisbon, Portugal (2004)
2. Levenstein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In Soviet Physics - Doklady 10, 707–710 (1966)
3. 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
4. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Working Notes of CLEF 2007 (2007)
5. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the Answer Validation Exercise 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
6. Rodrigo, Á., Peñas, A., Herrera, J., Verdejo, F.: Experiments of UNED at the Third Recognising Textual Entailment Challenge. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, pp. 89–94 (2007)
7. Rodrigo, Á., Peñas, A., Herrera, J., Verdejo, F.: The Effect of Entity Recognition on Answer Validation. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)

Adapting QA Components to Mine Answers in Speech Transcripts^{*}

Günter Neumann¹ and Rui Wang²

¹ LT lab, DFKI, Saarbrücken, Germany
neumann@dfki.de

² Saarland University, Saarbrücken, Germany
rwang@coli.uni-sb.de

Abstract. The paper describes QAST-v1 a robust question answering system for answering factoid questions in manual and automatic transcriptions of speech. The system is an adaptation of our text-based crosslingual open-domain QA system that we used for the CLEF main tasks.

1 Introduction

The focus of the new Question Answering on Speech Transcripts (QAsT) track within CLEF 2007 is on extracting answers to written factoid questions in manual and automatic transcripts of records of spoken lectures and meetings. Although the basic functionality of a QAsT-based system is similar to that of a textual QA-system the nature of the different scenarios and answer sources provoke new challenges.

The answer sources for CLEF and TREC-like systems are usually text documents like news articles or articles from Wikipedia. In general, an article of such a corpora describes a single topic using a linguistically and stylistically well-formed short text which has been created through a number of revision loops. In this sense, such an article can be considered as being created off-line for the prospective reader. By contrast, transcripts from lectures or meetings are live records of spontaneous speech produced incrementally or on-line in human-human interactions. Here, revisions (of errors or refinements) of utterances take place explicitly and immediately or not at all. Thus, speech transcripts also have to encode such properties of incremental language production, like word repetition, error corrections, refinements or interruptions. Consequently, transcripts are less well-formed, stylistic and fluent as written texts. Furthermore, in case of automatic transcripts errors and language gaps caused by the used automatic speech recognition system also make things not easier for a QAsT-based system.

^{*} The work presented here has been partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project HyLaP (FKZ: 01 IW F02) and by the EU funded project QALL-ME (FP6 IST-033860).

It seems that QA on speech transcripts demands a high degree of robustness and flexibility from the QA components and its architecture.

Nevertheless, the component architecture of a QAS_t-based system is similar to that of a textual QA-system and consists of the following core functionality: NL question analysis, retrieval of relevant snippets from speech transcripts, answer extraction, and answer selection. Therefore, we decided to develop our initial prototype QAS_t-v1 following the same underlying design principles that we used for our textual QA system and by the adaptation of some of its core components, cf. [3,4].

2 System Overview

The current information flow is as follows: In an off-line phase we firstly generate an inverted index for the speech corpora such that each sentence is considered as a single document and indexed by its word forms and named entities. In the question answering phase, a list of NL questions is passed to the system. Each NL question is analyzed by the named entity recognizer and by the question analysis component. The main output is a question object which represents the expected answer type (EAT) of the question and its relevant keywords. For example, the EAT of the question “Where is Southern Methodist University?” is LOCATION and the relevant keywords are “Southern Methodist University”. From the question object an IR-query expression is created in order to access the indexed document space. The IR-query for the example question is $\{+neTypes:LOCATION \text{ AND } +\text{“southern methodist university”}\}$ which can be read as “select only documents (in our case only sentences) which contain at least one location entity and the phrase Southern Methodist University”. In the answer extraction step all found location names are considered as answer candidates and the most frequent answer candidates are selected as answers to the question, e.g., “Dallas” and “Texas” are found as possible answers in the manual transcript of the lecture corpus. For each question a list of its N-best answers is returned. In the next sub-sections, we describe some of the core components in more detail.

2.1 Named Entity Recognition

Named Entity Recognition (NER) plays a central role in a factual QA architecture: Named entities are the answers of factual questions and as such define the range for the expected answer types. The answer types directly corresponds to the type of named entities.

There exists already a number NER components, but with different coverage of types. For that reason, we developed a hybrid NER approach where we combined three different NER components:

- LingPipe¹: It mainly covers PERSON, LOCATION, and ORGANIZATION names for English and co-references between pronouns and corresponding named entities. It realizes a supervised statistical based approach to NER.

¹ <http://www.alias-i.com/lingpipe/>

- Opennlp²: Its name finder is also based on a supervised statistical approach and covers mainly seven types of NEs for English, viz. PERSON, LOCATION, ORGANIZATION, DATE, TIME, MONEY, and PERCENTAGE.
- BiQueNER developed by our group. It is based on the semi-supervised approach developed by [1] and handles the following NE types: LANGUAGE, SYSTEM/METHOD, MEASURE, COLOUR, SHAPE, and MATERIAL.

All three NERs run in parallel on an input text. The individual results are combined via the IR-query construction process and the answer extraction process. In this way, also conflicting cases are handled like different NE readings and (implicit) partial or overlapping annotations.

2.2 Document Preprocessing

A sentence-oriented preprocessing determining only sentence boundaries, named entities (NE) and their co-references turned out to be a useful level of offline annotation of written texts, at least for the CLEF-kind of factual questions, cf. [3] for a detailed discussion. For that reason we decided to apply the same off-line preprocessing approach also to the QAsT collections. In particular the following steps are performed: 1) Extracting lines of words from the automatic speech transcripts so that both the manual and automatic transcript are in the same format. 2) Identification of sentence boundaries using the sentence splitter of the Opennlp tool which is based on maximum entropy modeling. We are currently using the language model the sentence splitter comes with which is optimized for written texts. 3) Annotation of the sentences with recognized named entities.

The preprocessed documents are further processed by the IR-development engine Lucene, cf. [2]. We are using Lucene in such a way that for all extracted named entities and content words, Lucene provides indexes which point to the corresponding sentences directly. Especially in the case of named entities type-based indexes are created which support the specification of type constraints in an IR-query. This will not only narrow the amount of data being analyzed for answer extraction, but will also guarantee the existence of an answer candidate.

2.3 Question Processing and Sentence Retrieval

In the current QAsT 2007 task setting natural language questions are specified in written form. For this reason we were able to integrate the question parser from our textual QA-system into QAsT-v1. The question parser computes for each question a syntactic dependency tree (which also contains recognized named entities) and semantic information like question type, the expected answer type, and the question focus, cf. [3] for details.

In a second step the result of the question parser is mapped to an ordered set of alternative IR-queries following the same approach as in our textual QA system, cf. [3].

² <http://opennlp.sourceforge.net/>

3 Results and Discussion

We took part in the tasks:

- T1: Question-Answering in manual transcriptions of lectures;
- T2: Question-Answering in automatic transcriptions of lectures;

In both cases the CHIL corpus was used which was adapted by the organizers for the QAsT 2007 track. It consists of around 25 hours (around 1 hour per lecture) both manually and automatically transcribed. The language is European English, mostly spoken by non-native speakers.

We submitted only one run to each task and the table below shows the results we obtained:

Run	task	Questions returned (#) [98]	Correct answers (#)	MRR	Accuracy
dfki1.t1	T1	98	19	0.17	0.15
dfki1.t2	T2	98	9	0.09	0.09

where MRR is the Mean Reciprocal Rank that measures how well ranked is the right answer in the list of 5 possible answers in average. Accuracy is the fraction of correct answers ranked in the first position in the list of 5 possible answers.

The currently low number of returned correct answers has two main error sources. On the one hand side, the coverage and quality of the named entity recognizers are low. This is probably due to the fact that we used the languages models that were created from written texts. One possible solution is to improve the corpus preprocessing step, especially the sentence splitter and the repairment of errors like word repetition. Another possible source of improvement is the development of annotated training corpus of speech transcripts for named entities. Both activities surely demand further research and resources.

On the other hand side, the performance of the answer extraction process strongly depends on the coverage and quality of the question analysis tool. We will improve this by extending the current coverage of the English Wh-grammar, especially by extending the mapping of general verbs and nouns to corresponding expected answer types and by exploiting strategies that validate the semantic type consistency between the relevant nouns and verbs of a question.

References

1. Collins, M., Singer, Y.: Unsupervised models for named entity classification (1999)
2. Hatcher, E., Gospodnetic, O.: Lucene in Action. In Action series. Manning Publications Co., Greenwich (2004)
3. Neumann, G., Sacaleanu, S.: Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 411–422. Springer, Heidelberg (2005)
4. Sacaleanu, B., Neumann, G.: Dfki-lt at the CLEF 2006 multiple language question answering track. In: Working notes of CLEF 2006, Alicante, Spain (August 2006)

The LIMSI Participation in the QAst Track

Sophie Rosset, Olivier Galibert, Gilles Adda, and Eric Bilinski

Spoken Language Processing Group, LIMSI-CNRS,
B.P. 133, 91403 Orsay cedex, France
`{firstname.lastname}@limsi.fr`

Abstract. In this paper, we present two different question-answering systems on speech transcripts which participated to the QAst 2007 evaluation. These two systems are based on a complete and multi-level analysis of both queries and documents. The first system uses handcrafted rules for small text fragments (snippet) selection and answer extraction. The second one replaces the handcrafting with an automatically generated research descriptor. A score based on those descriptors is used to select documents and snippets. The extraction and scoring of candidate answers is based on proximity measurements within the research descriptor elements and a number of secondary factors. The evaluation results are ranged from 17% to 39% as accuracy depending on the tasks.

1 Introduction

In the QA and Information Retrieval domains progress has been demonstrated via evaluation campaigns for both open and limited domains [1,2,3]. In these evaluations, independant questions are presented to the systems which has to provide one answer extracted from textual data to each question. Recently, there has been growing interest in extracting information from multimedia data such as meetings, lectures... Spoken data is different from textual data in various ways. The grammatical structure of spontaneous speech is quite different from written discourse and include various types of disfluencies. The lecture and interactive meeting data provided in QAst evaluation [4] are particularly difficult due to run-on sentences and interruptions. Typical textual QA systems are composed of question analysis, information retrieval and answer extraction components [1,5]. The answer extraction component is quite complex and involves natural language analysis, pattern matching and sometimes even logical inference [6]. Most of these natural language tools are not designed to handle spoken phenomena.

In this paper, we present the architecture of the two QA systems developed in LIMSI for the QAst evaluation. Our QA systems are part of an interactive and bilingual (English and French) QA system called Ritel [7] which specifically addresses speed issues. The following sections present the documents and queries pre-processing and the non-contextual analysis which are common to both systems. The section 3 describes the previous system (System 1). Section 4 presents the new system (System 2). Section 5 finally presents the results for these two systems on both development and test data.

2 Analysis of Documents and Queries

Usually, the syntactic/semantic analysis is different for the document and for the query; our approach is to perform the same complete and multilevel analysis on both queries and documents. There are several reasons for this. First of all, the system has to deal with both transcribed speech (transcriptions of meetings and lectures, user utterances) and text documents, so there should be a common analysis that takes into account the specificities of both data types. Moreover, incorrect analysis due to the lack of context or limitations of hand-coded rules are likely to happen on both data types, so using the same strategy for document and utterance analysis helps to reduce their negative impact. In order to use the same analysis module for all kinds of data, we should transform the query and the documents, which may come from different modality (text, manual transcripts, automatic transcripts) in order to have a common representation of the sentence, word, etc. This process is the normalization.

2.1 Normalization

Normalization, in our application, is the process by which *raw* texts are converted to a text form where words and numbers are unambiguously delimited, punctuation is separated from words, and the text is split into sentence-like segments (or as close to sentences as is reasonably possible). Different normalization steps are applied, depending of the kind of input data; these steps are:

1. Separating words and numbers from punctuation (for AMI data, T3 and T4 tasks)
2. Reconstructing correct case for the words (for AMI data, T3 and T4 tasks).
3. Adding punctuation (for AMI and CHIL data, T2 and T4 tasks).
4. Splitting into sentences at period marks (for all tasks).

Reconstructing the case and adding punctuation is done in the same process based on using a fully-cased, punctuated language model [8]. A word graph was built covering all the possible variants (all possible punctuations added between words, all possible word cases), and a 4-gram language model was used to select the most probable hypothesis. The language model was estimated on House of Commons Daily Debates, final edition of the European Parliament Proceedings and various newspapers archives. The final result, with uppercase only on proper nouns and words clearly separated by white-spaces, is then passed to the non-contextual analysis.

2.2 Non Contextual Analysis Module

The analysis is considered *non-contextual* because each sentence is processed in isolation. The general objective of this analysis is to find the bits of information that may be of use for search and extraction, which we call *pertinent information chunks*. These belong to different categories: named entities, linguistic entities

(e.g. verbs, prepositions), or specific entities (e.g. scores). All words that do not fall into such chunks are automatically grouped into chunks via a longest-match strategy. In the following sections, the types of entities handled by the system are described, along with how they are recognized.

Definition of Entities. Following commonly adopted definitions, the named entities are expressions that denote locations, people, companies, times, and monetary amounts. These entities have commonly known and accepted names. For example if the country France is a named entity, “capital of France” is not a named entity. However our experience is that the information present in the named entities is not sufficient to analyze the wide range of user utterances that may be found in lectures or meetings transcripts. Therefore we defined a set of specific entities in order to collect all observed information expressions contained in a corpus questions and texts from a variety of sources (proceedings, transcripts of lectures, dialogs etc.). Figure 1 summarizes the different entity types that are used.

Type of entities	Examples
<i>classical named entities</i>	pers: Romano Prodi ; Winston Churchill prod: Pulp Fiction ; Titanic time: third century ; 1998 ; June 30th org: European Commission ; NATO loc: Cambridge ; England
<i>extended named entities</i>	method: HMM, Gaussian mixture model event: the 9th conference on speech communication and technology amount: 500 ; two hundred and fifty thousand measure: year ; mile ; Hertz color red, spring green
<i>question markers</i>	Qpers: who wrote... ; who directed Titanic Qloc: where is IBM Qmeasure: what is the weight of the blue spoon headset
<i>linguistic chunk</i>	compound: language processing ; information technology verb: Roberto Martinez now knows the full size of the task adj_comp: the microphones would be similar to ... adj_sup: the biggest producer of cocoa of the world

Fig. 1. Examples of the main entity types

Automatic Detection of Typed Entities. The types we need to detect correspond to two levels of analysis: named-entity recognition and chunk-based shallow parsing. Various strategies for named-entity recognition using machine learning techniques have been proposed [9,10,11]. In these approaches, a statistically pertinent coverage of all defined types and subtypes induced the need of a large number of occurrences, and therefore rely on the availability of large annotated corpora which are difficult to build. Rule-based approaches to named-entity recognition (e.g. [12]) rely on morphosyntactic and/or syntactic analysis of the documents. However, in the present work, performing this sort of analysis is not feasible: the speech transcriptions are too noisy to allow for both accurate and robust linguistic analysis based on typical rules, and the processing time of

most of existing linguistic analyzers is not compatible with the high speed we require.

We decided to tackle the problem with rules based on regular expressions on words as in other works [13]: we allow the use of lists for initial detection, and the definition of local contexts and simple categorizations. The tool used to implement the rule-based automatic annotation system is called *Wmatch*. This engine matches (and substitutes) regular expressions using words as the base unit instead of characters. This property allows for a more readable syntax than traditional regular expressions and enables the use of classes (lists of words) and macros (sub-expressions in-line in a larger expression). *Wmatch* includes also NLP-oriented features like strategies for prioritizing rule application, recursive substitution modes, word tagging (for tags like noun, verb...), word categories (number, acronym, proper name...). It has multiple input and output formats, including an XML-based one for interoperability and to allow chaining of instances of the tool with different rule sets. Rules are pre-analyzed and optimized in several ways, and stored in compact format in order to speed up the process. Analysis is multi-pass, and subsequent rule applications operate on the results of previous rule applications which can be enriched or modified. The full analysis comprises some 50 steps and takes roughly 4 ms on a typical user utterance (or document sentence). The analysis provides 96 different types of entities. Figure 2 shows an example of the analysis on a query (top) and on a transcription (bottom).

<code><_Qorg> which organization </> <_verb> provided </> <_det> a </> <_NN> significant amount </> <_prep> of </> <_NN> training data </> <_punct> ? </></code>
<code><_pro> it </> <_verb> 's </> <_adv> just </> <_prep_comp> sort of </> <_det> a </> <_NN> very pale </> <_color> blue </> <_conj> and </> <_det> a </> <_adj> light-up </> <_color> yellow </> <_punct> . </></code>

Fig. 2. Example annotation of a query: *which organization provided a significant amount of training data?* (top) and of a transcription *it's just sort of a very pale blue* (bottom)

3 Question-Answering System 1

The *Question-Answering* system handles search in documents of any types (news articles, web documents, transcribed broadcast news, etc.). For speed reasons, the documents are all available locally and preprocessed: they are first normalized, and then analyzed with the non-contextual analysis module. The (type, values) pairs are then managed by a specialized indexer for quick search and retrieval.

This somewhat bag-of-typed-words system [7] works in three steps:

1. Document query list creation: the entities found in the question are used to select a document query and an ordered list of back-off queries from a predefined handcrafted set. These queries are obtained by relaxing some of the

constraints on the presence of the entities, using a relative importance ordering (Named entity > NN > adj_comp > action > subs ...)

2. Snippet retrieval: we submit each query, according to their rank, to the indexation server, and stop as soon as we get document snippets (sentence or small groups of consecutive sentences) back.

3. Answer extraction and selection: the detection of the answer type has been extracted beforehand from the question, using Question Marker, Named, Non-specific and Extended Entities co-occurrences ($\_Qwho \rightarrow \_pers$ or $\_pers\_def$ or $\_org$). Therefore, we select the entities in the snippets with the expected type of the answer. At last, a clustering of the candidate answers is done, based on frequencies. The most frequent answer wins, and the distribution of the counts gives an idea of the confidence of the system in the answer.

4 Question-Answering System 2

System 1 has three main problems. First, the back-off queries lists require a large amount of maintenance work and will never cover all of the combinations of entities which may be found in the questions. Second, the answer selection uses only frequencies of occurrence, often ending up with lists of first-rank candidate answers with the same score. And finally, The system answering speed directly depends on the number of snippets to retrieve which may sometimes be very large.

A new system (System 2) has been designed to solve these problems. We have kept the three steps described in section 3, with some major changes. In step 1, instead of instantiating document queries from a large number of preexisting handcrafted rules (about 5000), we generate a research descriptor using a very small set of rules (about 10). In step 2, a score is calculated from the proximity between the research descriptor and the document and snippets, in order to choose the most relevant ones. In step 3, the answer is selected according to a score which takes into account many different features and tuning parameters, which allow an automatic and efficient adaptation.

4.1 Research Descriptor Generation

The first step of System 2 is to build a research descriptor (data descriptor record, DDR) which contains the important elements of the question, and the possible answer types with associated weight. Some elements are marked as *critical*, which makes them mandatory in future steps, while others are *secondary*. The element extraction and weighting is based on a empirical classification of the element types in importance levels. Answer types are predicted through rules based on combinations of elements of the question. The Figure 3 shows an example of a DDR.

4.2 Documents and Snippets Selection and Scoring

Each of the document is scored with geometric mean of the number of occurrences of all the DDR elements which appear in it. Using a geometric mean

```

question: in which company Bart works as a project manager ?
ddr:
{ w=1, critical, pers, Bart},
{ w=1, critical, NN, project manager },
{ w=1, secondary, action, works },
answer_type = {
  { w=1.0, type=orgof },
  { w=1.0, type=organisation },
  { w=0.3, type=loc },
  { w=0.1, type=acronym },
  { w=0.1, type=np },
}

```

Fig. 3. Example of a DDR constructed from the question *in which company Bart works as a project manager*; each element contains a weight w , their importance for future steps, and the pair (type,value); each possible answer type contains a weight w and the type of the answer

prevents from rescaling problems due to some elements being naturally more frequent. The documents are sorted by score and the n -best ones are kept. The speed of the entire system can be controlled by choosing n , the whole system being in practice io-bound rather than cpu-bound.

The selected documents are then loaded and all the lines in a predefined window (2-10 lines depending on question types) from the critical elements are kept, creating snippets. Each snippet is scored using the geometrical mean of the number of occurrences of all the DDR elements which appear in the snippet, smoothed with the document score.

4.3 Answer Extraction, Scoring and Clustering

In each snippet all the elements which type is one of the predicted possible answer types are candidate answers. We associate to each candidate answer A a score $S(A)$:

$$S(A) = \frac{[w(A) \sum_E \max_{e=E} \frac{w(E)}{(1+d(e,A))^\alpha}]^{1-\gamma} \times S_{snip}^\gamma}{C_d(A)^\beta C_s(A)^\delta} \quad (1)$$

In which:

- $d(e, A)$ is the distance to each element e of the snippet, instantiating a search element E of the DDR
- $C_s(A)$ is the number of occurrences of A in the extracted snippets, $C_d(A)$ in the whole document collection
- S_{snip} is the extracted snippet score (see 4.2)
- $w(A)$ is the weight of the answer type and $w(E)$ the weight of the element E in the DDR

• α, β, γ and δ are tuning parameters estimated by systematic trials on the development data. $\alpha, \beta, \gamma \in [0, 1]$ and $\delta \in [-1, 1]$

An intuitive explanation of the formula is that each element of the DDR adds to the score of the candidate ($\sum_E$) proportionally to its weight ($w(E)$) and inversely proportionally to its distance of the candidate ($d(e, A)$). If multiple instance of the element are found in the snippet only the best one is kept ($\max_{e=E}$). The score is then smoothed with the snippet score (S_{snip}) and compensated in part with the candidate frequency in all the documents (C_d) and in the snippets (C_s).

The scores for identical (type,value) pairs are added together and give the final scoring for all the possible candidate answers.

5 Evaluation

The QAsT evaluation proposed 4 tasks : QA on manually transcribed seminar data (T1), QA on the same data automatically transcribed (T2), QA on manually transcribed meeting data (T3) and QA on the same data automatically transcribed (T4) (see [4] for details). T1 and T2 tasks were composed of an identical set of 98 questions; T3 task was composed of a different set of 96 questions and T4 task of a subset of 93 questions. Table 1 show the overall results with the 3 measures used in this evaluation. We submitted two runs, one for each system, for each of the four tasks. As required by the evaluation procedure, a maximum of 5 answers per question was provided.

Table 1. General Results. *Sys1* System 1; *Sys2* System 2; *Acc.* is the accuracy, *MRR* is the Mean Reciprocal Rank and *Recall* the total number of correct answers in the 5 returned answers. Between parenthesis are the results on the development data.

Task	System	Acc.	MRR	Recall
T1	Sys1	32.6% (74%)	0.37	43.8%
T1	Sys2	39.7% (94%)	0.46	57.1%
T2	Sys1	20.4% (24%)	0.23	28.5%
T2	Sys2	21.4% (34%)	0.24	28.5%
T3	Sys1	26.0% (28%)	0.28	32.2%
T3	Sys2	26.0% (72%)	0.31	41.6%
T4	Sys1	18.3% (20%)	0.19	22.6%
T4	Sys2	17.2% (32%)	0.19	22.6%

System 2 gets better results than System 1. The improvement of the Recall (9-13%) observed on T1, and T3 tasks for System 2 illustrates that automatic generation of document/snippet queries greatly improves the coverage as compared to handcrafted rules. System 2 did not perform better than System 1 on the T2 task. Further analysis is needed to understand why. In particular, there are major differences between the results on development and test data.

One of the critical point of the analysis, is the routing of the question in which we determine a rough class for the type of the answer (*language, location, ...*). The

Table 2. Routing evaluation. *All*: all questions; *LAN*: language; *LOC*: location; *MEA*: measure; *MET*: method/system; *ORG*: organization; *PER*: person; *TIM*: time; *SHAP*: shape; *COL*: colour. *Cor.*: correct routing. *Quest*: number of questions. Between parenthesis are the results on the development data.

		All	LAN	LOC	MEA	MET	ORG
T1	Cor.	72% (100%)	100% (100%)	89% (100%)	75% (100%)	17% (100%)	95% (100%)
	Quest.	98 (50)	4 (3)	9 (14)	28 (6)	18 (3)	20 (7)
T3	Cor.	80% (90%)	100% (-)	93% (100%)	83% (83%)	-	85% (71%)
	Quest.	96 (50)	2 (-)	14 (4)	12 (12)	-	13 (7)

		PER	TIM	SHA	COL	MAT
T1	Cor.	89% (100%)	80% (100%)	-	-	-
	Quest.	9 (12)	10 (5)	-	-	-
T3	Cor.	80% (88%)	71% (100%)	89% (100%)	73% (100%)	50% (100%)
	Quest.	15 (10)	14 (5)	9 (4)	11 (5)	6 (4)

Table 3. Results for Passage Retrieval for System 2. *Passage 5* the maximum of passage number is 5; *Passage without limit* there is no limit for the passage number; *Acc.* is the accuracy, *MRR* is the Mean Reciprocal Rank and *Recall* the total number of correct answers in the returned passages.

		Passage limit = 5			Passage without limit		
Task		Acc.	MRR	Recall	Acc.	MRR	Recall
T1		44.9%	0.52	67.3%	44.9%	0.53	71.4%
T2		29.6%	0.36	46.9%	29.6%	0.37	57.0%
T3		30.2%	0.37	47.9%	30.2%	0.38	68.8%
T4		18.3%	0.22	31.2%	18.3%	0.24	51.6%

results of the routing component are given in Table 2 with details by answer category. Two questions of T1/T2 and three of T3/T4 were not routed. We observed large differences with the results obtained on the development data, in particular with the *method*, *color* and *time* categories. One observation can be made: in the dev data all the three questions about method/systems contained the words *method* or *system* and none of all the 18 test questions. The analysis module has been built on corpus observations and seems too dependent on the development data. That can explain the absence of major differences between System 1 and System 2 for the T1/T2 tasks. Most of the wrongly routed questions have been routed to the generic answer type class. In System 1 this class selects specific entities (*method*, *models*, *system*, *language*...) over the other entity types for the possible answers. In System 2 no such adaptation to the task has been done and all possible entity types have equal priority.

The passage retrieval for System 2 may easily be evaluated. The Table 3 give the results on the passage retrieval in two conditions: with a limitation of the

number of passages at 5 and without limitation. The difference between the Recall on the snippets (how often the answer is present in the selected snippets) and the QA Accuracy shows that the extraction and the scoring of the answer have a reasonable margin for improvement. The difference between the snippet Recall and its Accuracy (about 30% for the no limit condition) illustrates that the snippet scoring can be also improved.

6 Conclusion and Future Work

We presented the Question Answering systems used for our participation to the QAsT evaluation. Two different systems have been used for this participation. The two main changes between System 1 and System 2 are the replacement of the large set of hand made rules by the automatic generation of a research descriptor, and the addition of an efficient scoring of the candidate answers. The results show that the System 2 outperforms the System 1. The main reasons are:

1. Better genericity through the use of a kind of expert system to generate the research descriptors.
2. More pertinent answer scoring using proximities which allows a smoothing of the results.
3. Presence of various tuning parameters which enable the adaption of the system to the various question and document types.

These systems have been evaluated on different data corresponding to different tasks. On the manually transcribed lectures, the best result is 39% for Accuracy, on manually transcribed meetings, 24% for Accuracy. There was no specific effort done on the automatically transcribed lectures and meetings, so the performances only give an idea of what can be done without trying to handle speech recognition errors. The best result is 18.3% on meeting and 21.3% on lectures. From the analysis presented in the previous section, performance can be improved at every step. For example, the analysis and routing component can be improved in order to better take into account some type of questions which should improve the answer typing and extraction. The scoring of the snippets and the candidate answers can also be improved. In particular some tuning parameters (as the weight of the transformations generated in the DDR) have not been optimized yet.

Acknowledgments

This work was partially funded by the European Commission under the FP6 Integrated Project IP 506909 CHIL and the LIMSI AI/ASP RITEL grant.

References

1. Voorhees, E.M., Buckland, L.P.: In: Voorhees, Buckland (eds.) *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)* (2006)
2. Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary (September 2007)
3. Ayache, C., Grau, B., Vilnat, A.: Evaluation of question-answering systems: The French EQueR-EVALDA Evaluation Campaign. In: *Proceedings of LREC 2006*, Genoa, Italy (2006)
4. Turmo, J., Comas, P., Ayache, C., Mostefa, D., Rosset, S., Lamel, L.: Overview of QAST 2007. In: *Working Notes of CLEF 2007 Workshop*, Budapest, Hungary (September 2007)
5. Harabagiu, S., Moldovan, D.: Question-Answering. In: Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford (2003)
6. Harabagiu, S., Hickl, A.: Methods for using textual entailment in Open-Domain question-answering. In: *Proceedings of COLING 2006*, Sydney, Australia (July 2006)
7. van Schooten, B., Rosset, S., Galibert, O., Max, A., op den Akker, R., Illouz, G.: Handling speech input in the Ritel QA dialogue system. In: *Proceedings of Interspeech 2007*, Antwerp, Belgium (August 2007)
8. Déchelotte, D., Schwenk, H., Adda, G., Gauvain, J.-L.: Improved Machine Translation of Speech-to-Text outputs. In: *Proceedings of Interspeech 2007*, Antwerp, Belgium (August 2007)
9. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: *Proceedings of ANLP 1997*, Washington, USA (1997)
10. Isozaki, H., Kazawa, H.: Efficient Support Vector Classifiers for Named Entity Recognition. In: *Proceedings of COLING*, Taipei (2002)
11. Surdeanu, M., Turmo, J., Comelles, E.: Named Entity Recognition from spontaneous Open-Domain Speech. In: *Proceedings of InterSpeech 2005*, Lisbon, Portugal (2005)
12. Wolinski, F., Vichot, F., Dillet, B.: Automatic Processing of Proper Names in Texts. In: *Proceedings of EACL 1995*, Dublin, Ireland (1995)
13. Sekine, S.: Definition, dictionaries and tagger of Extended Named Entity hierarchy. In: *Proceedings of LREC 2004*, Lisbon, Portugal (2004)

Robust Question Answering for Speech Transcripts Using Minimal Syntactic Analysis

Pere R. Comas¹, Jordi Turmo¹, and Mihai Surdeanu²

¹ TALP Research Center, Technical University of Catalonia (UPC),
pcomas@lsi.upc.edu, turmo@lsi.upc.edu

² Barcelona Media Innovation Center
mihai.surdeanu@barcelonamedia.org

Abstract. This paper describes the participation of the Technical University of Catalonia in the CLEF 2007 Question Answering on Speech Transcripts track. For the processing of manual transcripts we have deployed a robust factual Question Answering that uses minimal syntactic information. For the handling of automatic transcripts we combine the QA system with a novel Passage Retrieval and Answer Extraction engine, which is based on a sequence alignment algorithm that searches for “sounds like” sequences in the document collection. We have also enriched the NERC with phonetic features to facilitate the recognition of named entities even when they are incorrectly transcribed.

Keyword: Question Answering, Spoken Document Retrieval, Phonetic Distance.

1 Introduction

The CLEF 2007 Question Answering on Speech Transcripts (QAsT) track [8] consists of the following four tasks: Question Answering (QA) in manual transcripts of recorded lectures (T1) and their corresponding automatic transcripts (T2), and QA in manual transcripts of recorded meetings (T3) and their corresponding automatic transcripts (T4).

For tasks T1 and T3 we have adapted a QA system and Named Entity Recognizer and Classifier (NERC) that we previously developed for manual speech transcripts [6,7]. For the handling of automatic transcripts (T2 and T4) we implemented two significant changes: (a) for Passage Retrieval and Answer Extraction we designed a novel keyword matching engine that relies on phonetical similarity (instead of string match) to overcome the errors introduced by the ASR, and (b) we enriched the NERC with phonetic features to facilitate the recognition of named entities (NEs) even when they are incorrectly transcribed.

2 Overview of the System Architecture

The architecture of our QA system follows a commonly-used schema which splits the process into three phases performed sequentially: Question Processing (QP), Passage Retrieval (PR), and Answer Extraction (AE).

1M: *"The pattern frequency relevance rate indicates the ratio of relevant documents"*
1A: *"the putt and frequency illustrating the case the ratio of relevant documents"*
2M: *"The host system it is a UNIX Sun workstation"*
2A: *"that of system it is a unique set some workstation"*

Fig. 1. Examples of manual (M) and automatic (A) transcripts

2.1 QA System for Manual Transcripts

For the processing of manual transcripts we used an improved version of our system introduced in [6]. We describe it briefly below.

QP: The main goal of this component is to detect the type of the expected answer. We currently recognize the 53 open-domain answer types from [4] plus 3 types specific to QAs corpora (i.e., **system/method**, **shape**, and **material**). The answer types are extracted using a multi-class Perceptron classifier and a rich set of lexical, semantic and syntactic features. This classifier obtains an accuracy of 88.5% on the corpus of [4]. Additionally, the QP component extracts and ranks relevant keywords from the question

PR: This component retrieves a set of relevant passages from the document collection, given the previously extracted question keywords. The PR algorithm uses a query relaxation procedure that iteratively adjusts the number of keywords used for retrieval and their proximity until the quality of the recovered information is satisfactory (see [6]). In each iteration a Document Retrieval application (Lucene IR engine) fetches the documents relevant for the current query and a subsequent passage construction module builds passages as segments where two consecutive keyword occurrences are separated by at most t words.

AE: Identifies the exact answer to the given question within the retrieved passages. First, answer candidates are identified as the set of NEs that occur in these passages and have the same type as the answer type detected by QP. Then, these candidates are ranked using a scoring function based on a set of heuristics that measure keyword distance and density[5].

2.2 QA System for Automatic Transcripts

The state of the art in ASR technology is far from perfect. For example, the word error rate (WER) of the meetings automatic transcripts is around 38% and the WER of the lectures is over 20%. Figure 1 shows two real examples of common errors when generating automatic transcripts. From the point of view of QA, imperfect transcripts create the following problems: (a) The keywords identified as relevant by QP define the context where the correct answer appears. They are used for PR and AE. When these specific keywords are incorrectly transcribed by the ASR, all these tasks are in jeopardy. (b) Most NEs (candidate answers) appear as proper nouns with low frequency in the corpora. Due to this low frequency it is unlikely that the ASR language models include them. Then

it is probable that ASR incorrectly recognizes the NEs relevant for the AE component.

To address these issues specific to automatically-generated transcripts we have developed a novel QA system by changing the PR, AE and NERC components. The main difference between the new PR and AE modules and those used to process manual transcripts is the strategy for keyword searching. Our hypothesis is that an approximated matching between the automatic transcripts and the question keywords, according phonetic similarity can perform better than classical IR techniques for written text. automatic transcripts and question keywords extracted by QP are deterministically transformed to phonetic sequences. Then we use a novel retrieval engine named PHAST, which computes document (or passage or answer context) relevance based on approximated matching of phonetic sequences. PHAST is detailed in Section 4.

3 Named Entity Recognition and Classification

As described before, we extract candidate answers from the NEs that occur in the passages retrieved by the PR component. We detail below the strategies used for NERC in both manual and automatic transcripts.

NERC for Manual Transcripts. Our initial idea was to use the NERC we developed previously for the processing of speech transcripts [7]. One change from the previous system is that we replaced the existing SVM classifiers with a multi-class Perceptron. To verify the validity of this approach we annotated the NEs that occur in the QAsT development corpus with their types (i.e., person, organization, location, language, measure, system/method and time) and used an 80–20% corpus split for training and testing for both lectures and meetings corpora. This experiment indicated that the development data is sufficient for good generalization for meetings (a F_1 score of +75 points in the development test partition) but it is insufficient in lectures: 33 points. This is most likely caused by the small size of the development corpus and the large number of topics addressed. To compensate for the insufficient training data we decided to perform a combination of several NERC models for this task. We merged the outputs of: (a) a rule-based NERC developed previously [6], (b) the NERC trained on the existing development data, and (c) the NERC trained on the CoNLL English corpus.¹ We used the above priority ordering for conflict resolution in case of overlapping assignments (e.g., lectures model has higher priority than the CoNLL model). After model combination the NERC F_1 score in the development test partition did not improve but the recall did increase, so we decided to use this combination strategy in the testing since recall is paramount for QA.

NERC for Automatic Transcripts. We used a similar framework for the processing of automatic transcripts: we annotated the development corpora and trained specific NERC models for lectures and meetings. The significant difference is that here we expand the classifiers' feature sets with phonetic attributes.

¹ <http://cnts.ua.ac.be/conll2002/ner>

These features are motivated by the fact that even when the ASR incorrectly transcribes NEs the phonetic structure is by and large maintained in the transcript (e.g. in Figure 1 the name “*Sun*” is recognized as “*some*”). We used an unsupervised hierarchical clustering algorithm that groups tokens based on the similarity of their phonetic sequences. The stop condition of the algorithm is set to reach a local maximum of the Calinski criterion [2]. Then the cluster of each token is added as a feature (e.g. “*Sun*” and “*some*” share the same cluster), which helps the NERC model generalize from the correct to the incorrect transcript. We also added phonetic features that model prefix and suffix similarity.

4 The Phonetic Sequence Alignment Algorithm

This section describes PHAST, the phonetic sequence alignment algorithm we used for keyword matching. The same algorithm can be used for PR and identification of answer contexts. PHAST is based on BLAST[1], an algorithm from the field of pattern matching in bioinformatics, which we adapted to work with phone sequences instead of protein sequences. In our case, the input data is a transcript collection D transformed to phonetic sequences and a set of query terms KW also mapped to phonetic sequences.

PHAST is detailed in Algorithm 1. The procedure works as follows: function *detection*() detects subsequences of transcript d at phone number r with moderate resemblance with keyword w , then *extension*() computes a similarity score s between d and w at r , and *relevant*() judges how this occurrence at r is relevant to term frequency. Function *detection*() uses a deterministic finite automaton (DFA) length n from w while scanning d . Given that the ill-transcribed words keep phonetic resemblance with the original words, our hypothesis is that short sequences of n phones will be in the original position. Function *extension*() is a measure of phonetic similarity [3]. We compute the similarity s of two sequences using the edit distance (Levenshtein distance) with a cost function that measures inter-phone similarity. The score s is a bounded non-integer value normalised into the interval $[0, 1]$ Function *relevant*() considers a *hit* any matching with the score above some fixed threshold. In the context of document retrieval,

Algorithm 1.

PHAST algorithm

Parameter: $\mathcal{D}$, collection of phonetically transcribed documents

Parameter: $\mathcal{KW}$, set of phonetically transcribed keywords

```

1: for all  $d \in \mathcal{D}, w \in \mathcal{KW}$  do
2:   while  $h = \text{detection}(w, d)$  do
3:      $s = \text{extension}(w, h)$ 
4:     if  $\text{relevant}(s, h)$  then
5:       mark  $w$  as matched  $\rightarrow$  update  $tf(w, d)$ 
6:     end if
7:   end while
8: end for
```

Automatic transcript: “that of system it is a unique set some workstation”

	jʊn	← <i>detection</i> _φ	
... ðæt ʌβ sistəm ɪt ɪz ə	jʊnik	sɛt sʌm	wəʊrksteɪʃən...
	jʊnik s	sʌn	← <i>extension</i> _φ

Fig. 2. Search of term “UNIX-Sun”

term frequency is computed by adding the scores of these hits. For PR and AE we used all relevant matchings in the algorithms described in Section 2.1. Figure 2 shows an example of how functions *detection* and *extension* are used. The sentence 2A from Figure 1 is transcribed to a sequence of phones. The query word w is the term “UNIX-Sun”, which is transcribed as [jʊniks sʌn].² Term w exists in the manual transcript 2M but not in the automatic transcript 2A. In the first step, *detection* finds the 3-gram [jʊn]. In the second step, *extension* extends it by matching the rest of [jʊniks sʌn] with the phones surrounding [jʊn] in the automatic transcript.

5 Experimental Results

UPC participated in all four QAst tasks. Initially, each task included 100 test questions, but a few ones were removed due to various problems. The final question distribution was: 98 questions in T1 and T2, 96 in T3, and 93 in T4. In the tasks T1 and T3 we submitted one run using the system described in Section 2.1 (QA_m). In the tasks based on automatic transcripts (T2 and T4) we submitted two runs: one using QA_m, and another using the system tailored for automatic transcripts as seen in Section 2.2 (QA_a). We report two measures: (a) TOP k , which assigns to a question a score of 1 only if the system provided a correct answer in the top k returned; and (b) Mean Reciprocal Rank (MRR), which assigns to a question a score of $1/k$, where k is the position of the correct answer, or 0 if no correct answer is found. An answer is considered correct by the human evaluators if it contains the complete answer and nothing more, and it is supported by the corresponding document. If an answer was incomplete or it included more information than necessary or the document did not provide the justification for the answer, the answer was considered incorrect.

The corpora were pre-processed as follows. We deleted word fragment markers, onomatopoeias, and utterance information in manual transcripts (tasks T1 and T3). Speaker turns in tasks T3 and T4 were substituted by sentence boundaries (this influences our answer ranking heuristics [6]) and the dialog was collapsed into a single document. For T2, all non-word tokens were deleted (e.g., “{breath}”), utterance markers and fragment words were eliminated. Then the documents were pre-processed by a POS tagger, lemmatizer, and NERC.

Table 1 summarizes our overall results. It shows that moving from manual transcripts to automatic transcripts (i.e., the difference of T1/T2, and T3/T4)

² We use the international phonetic alphabet (IPA): www.arts.gla.ac.uk/IPA/

Table 1. Overall results for the four QAs tasks. For task T3 we report scores using a post-deadline submission where some bugs in our output formatting script were fixed.

Task, System	#Q	MRR	TOP1	TOP5	Task, System	#Q	MRR	TOP1	TOP5
T1, QA _m	98	0.53	50	54	T3, QA _m	96	0.26	24	27
T2, QA _a	98	0.25	24	29	T4, QA _a	93	0.15	12	17
T2, QA _m	98	0.37	35	37	T4, QA _m	93	0.22	20	22

Table 2. Distribution of correct answers (TOP5) according to answer type. Org = organization, Per = person, Tim = time, Mea = measure, Met/Sys = method/system, Mat = material, Col = color.

Task and System	Org	Per	Loc	Tim	Mea	Met/Sys	Lan	Sha	Mat	Col
T1, QA _m	10/20	8/9	4/9	7/10	12/28	10/18	3/4	-	-	-
T2, QA _a	6/20	4/9	2/9	6/10	10/28	5/18	3/4	-	-	-
T2, QA _m	8/20	3/9	3/9	6/10	7/28	7/18	2/4	-	-	-
T3, QA _m	5/13	8/15	6/14	1/14	4/12	-	1/2	5/9	4/6	8/11
T4, QA _a	2/13	3/15	2/14	1/14	2/12	-	0/2	3/9	1/6	4/11
T4, QA _m	3/13	2/15	3/14	1/14	4/12	-	1/2	3/9	1/6	5/11

Table 3. Error analysis of the QA system components

Task and System	#Questions	QC Correct	PR Correct	QC&PR Correct	TOP1
T1, QA _m	98	67	82	54	50
T2, QA _a	98	67	80	29	24
T2, QA _m	98	67	76	37	36
T3, QA _m	96	87	73	66	25
T4, QA _a	93	87	52	47	13
T4, QA _m	93	87	58	53	21

yields a drop in TOP1 score of 29% in lectures and 16% in meetings. To our knowledge, this is the first time that such an analysis is performed for QA. It is encouraging to see that our scores are higher than the mean scores observed in TREC 2006 QA evaluation. Surprisingly, the performance drop is smaller for the meetings, even though these transcripts had a higher WER than lectures (38% versus 20%). The explanation is that, because the meetings tasks are harder due to the larger corpus and the more ambiguous question terms, we answer only the “easier” questions in the manual transcripts. Such questions tend to have a larger number of question keywords and answers that appear repeatedly in the collection, so the probability that the system encounter a valid answer even in automatic transcripts is large. In contrast, lecture corpus is very small, so one ASR mistake may be sufficient to lose the only existing correct answer for a given question. Based on these experiments, we can conclude that the QA performance drop follows the WER in small corpora with little redundancy and is smaller than WER in larger corpora with enough redundancy.

One unexpected result in this evaluation was that the QA_a system performed worse than the QA_m system on automatic transcripts (tasks T3 and T4), even though the QA_a system was designed for automatic transcripts. The explanation is two fold. First, with our current parameter setting, the PHAST algorithm triggered too many false keyword matches due to a relaxed approximated match. This yielded sets of candidate passages and answers with a lot of noise that was hard to filter out. Second, the NERC training data (i.e., the development corpus) was insufficient to learn correct phonetic generalizations, so many answer candidates were missed in automatic transcripts. Nevertheless, we believe that the architecture of the QA_a system is a good long-term investment because it is the only one of the two systems developed that can address the phenomena specific to automatic transcripts.

Table 2 shows the distribution of correct answers for all tasks according to the answer type. The table indicates that our system had a particularly hard time answering questions in task T3/T4, when the answer type was a NE of types **Org**, **Loc**, **Tim**, or **Mea**. These entity types have a high variation in the corpus and our NERC could not generalize well given the small amount of training data available. This suggests that a better strategy for NERC could be to train an open-domain NERC, where large annotated corpora are available, and use domain transfer techniques to adapt the open-domain system to this domain. The performance drop-off between manual and automatic transcripts is similar in all NE types.

Finally, table 3 summarizes the error analysis of QP, PR, and AE. The “QC Correct” column is the number of questions with the answer type correctly detected by QP. “PR Correct” is the number of questions where at least one passage with the correct answer was retrieved. “QC & PR Correct” is the number of questions where QP prediction is correct *and* PR retrieved a correct passage. We can draw several important observations from this error analysis: QP performs significantly worse for T1 question set than T3 question set. This suggests that in this evaluation T1 questions were more domain specific than T3 questions. Also, PR performs similarly to the state of the art for written text for tasks T1, T2, and T3, but it suffers an important performance hit on task T4, where we processed automatic transcripts with the highest WER (38%). This proves that PR is indeed affected by a high WER. PR using PHAST performed better than the PR with exact keyword match for task T2 and worse for task T4. As previously mentioned, this worse-than-expected behavior of PHAST was due to the many false-positive keyword matches generated in our current setup. We leave the better tuning of PHAST for the various QA tasks as future work. Finally, for tasks T1/T2, when the QA system reaches AE with the correct information (i.e., the “QC & PR Correct” in the table), AE performed very well: we answered most of those questions correctly on the first position. This indicates that both the NERC and the answer ranking performed well. For tasks T3/T4, the story is no longer the same: we suffer the biggest performance hit in AE. We manually inspected these errors and the conclusion was that in most of the cases the fault can be assigned to the NERC, which failed to recognize entity mentions

that were correct answers in both manual and automatic transcripts. This problem was mitigated in tasks T1/T2 with a combination of NERC models, which included a rule-based system previously developed for the lectures domain.

6 Conclusions

This paper describes UPC's participation in the CLEF 2007 Question Answering on Speech Transcripts track. We were one of the few participants that submitted runs in all four sub-tasks and we obtained the highest overall score. Our best performing runs have TOP1 scores that range from 0.21 (on automatic transcripts with WER of 38%) to 0.51 (on manual transcripts).

In this evaluation we analyzed the behavior of two systems differing in that one is tailored for manual transcripts while the other is tailored for automatic transcripts (uses approximate keyword search based on phonetic distances and a NERC enhanced with phonetic features). In all four tasks we obtained the best performance with the system designed for manual transcripts. This system performed better than expected on automatic transcripts for two reasons: first, it only requires the document collection to be POS tagged, and this technology is robust enough to function well on unperfect automatic transcripts. Second, the query relaxation algorithm adapts well to automatic transcripts: question terms that are incorrectly transcribed are automatically discarded. The system designed for automatic transcripts performed worse than expected because the approximated keyword match algorithm generated too many false-positive, introducing noise in the candidate sets of passages and answers, and also it was impossible for the NERC to detect the correct NEs in the new passages retrieved. Nevertheless, we believe that this approach is a good long-term research direction because it can truly address the phenomena specific to automatic transcripts.

Acknowledgements

This work has been partially funded by the European Commission (CHIL, IST-2004-506909) and the Spanish Ministry of Science (TEXTMESS project).

References

1. Altschul, S., Gish, W., Miller, W., Meyers, E.W., Lipman, D.J.: Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410 (1990)
2. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. *Communications in Statistics* 3 (1974)
3. Kondrak, G.: Algorithms for Language Reconstruction. PhD thesis, University of Toronto (2002)
4. Li, X., Roth, D.: Learning question classifiers: The role of semantic information. *Journal of Natural Language Engineering* (2005)

5. Paşca, M.: High-performance, open-domain question answering from large text collections. PhD thesis, Southern Methodist University, Dallas, TX (2001)
6. Surdeanu, M., Dominguez-Sal, D., Comas, P.R.: Design and performance analysis of a factoid question answering system for spontaneous speech transcriptions. In: Proceedings of the INTERSPEECH (2006)
7. Surdeanu, M., Turmo, J., Comelles, E.: Named entity recognition from spontaneous open-domain speech. In: Proceedings of the INTERSPEECH (2005)
8. Turmo, J., Comas, P.R., Ayache, C., Mostefa, D., Rosset, S., Lamel, L.: Overview of QAST 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 249–256. Springer, Heidelberg (2008)

Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task

Michael Grubinger¹, Paul Clough², Allan Hanbury³, and Henning Müller^{4,5}

¹ Victoria University, Melbourne, Australia

² Sheffield University, Sheffield, United Kingdom

³ Vienna University of Technology, Vienna, Austria

⁴ University and Hospitals of Geneva, Switzerland

⁵ University of Applied Sciences, Sierre, Switzerland

michael.grubinger@research.vu.edu.au

Abstract. The general photographic ad-hoc retrieval task of the *ImageCLEF 2007* evaluation campaign is described. This task provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval from generic photographic collections. In 2007, the evaluation objective concentrated on retrieval of lightly annotated images, a new challenge that attracted a large number of submissions: a total of 20 participating groups submitted 616 system runs. This paper summarises the components used in the benchmark, including the document collection and the search tasks, and presents an analysis of the submissions and the results.

1 Introduction

ImageCLEFphoto 2007 provides a system-centered evaluation for multilingual visual information retrieval from generic photographic collections (*i.e.* containing everyday real-world photographs akin to those that can frequently be found in private photographic collections). The evaluation scenario is similar to the classic TREC¹ ad-hoc retrieval task: simulation of the situation in which a system knows the set of documents to be searched, but cannot anticipate the particular topic that will be investigated (*i.e.* topics are not known to the system in advance) [1]. The goal of the simulation is: given an alphanumeric statement (and/or sample images) describing a user information need, find as many relevant images as possible from the given collection (with the query language either being identical or different from that used to describe the images).

The objective of *ImageCLEFphoto 2007* comprised the evaluation of multilingual visual information retrieval from a generic collection of lightly annotated photographs (*i.e.* containing only short captions such as the title, location, date or additional notes, but without a semantic description of the photograph). This new challenge allows for the investigation of the following research questions:

¹ <http://trec.nist.gov/>

- Are traditional text retrieval methods still applicable for such short captions?
- How significant is the choice of the retrieval language?
- How does the retrieval performance compare to retrieval from collections containing fully annotated images (*ImageCLEFphoto 2006*)?

One major goal of *ImageCLEFphoto 2007* was to attract more content-based image retrieval approaches, as most of the retrieval approaches in previous years had predominately been concept-based. The reduced alphanumeric semantic information provided with the image collection should support this goal as content-based retrieval techniques become more significant with increasingly reduced image captions.

2 Methods

Similar to *ImageCLEFphoto 2006* [2], we generated a subset of the *IAPR TC-12 Benchmark* to provide the evaluation resources for *ImageCLEFphoto 2007*. This section provides more information on these individual components: the document collection, the query topics, relevance judgments and performance indicators.

2.1 Document Collection

The document collection of *IAPR TC-12 Benchmark* contains 20,000 colour photos taken from locations around the world and comprises a varying cross-section of still natural images. More information on the design and implementation of test collection, created under *Technical Committee 12* (TC-12) of the *International Association of Pattern Recognition* (IAPR²), can be found in [3].



```
<DOC>
<DOCID>00000000000000000000000000000000</DOCID>
<TITLE>Flamingo</TITLE>
<DESCRIPTION>A group of flamingos standing in the water.
The flamingos are standing in the water, and the water is
shallow. The flamingos are standing in the water, and the
water is shallow. The flamingos are standing in the water,
and the water is shallow.
</DESCRIPTION>
<NOTES>A group of flamingos standing in the water. The
flamingos are standing in the water, and the water is
shallow. The flamingos are standing in the water, and the
water is shallow.
</NOTES>
<PROVIDER>IAPR</PROVIDER>
</DOC>
```

Fig. 1. Sample image caption

Each image in the collection has a corresponding semi-structured caption consisting of the following seven fields: (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

² <http://www.iapr.org/>

(6) where and (7) when the photo was taken. Figure 2.1 shows a sample image with its corresponding English annotation.

These annotations are stored in a database, allowing the creation of collection subsets with respect to a variety of particular parameters (*e.g.* which caption fields to use). Based on the feedback from participants of previous evaluation tasks, the following was provided for *ImageCLEFphoto 2007*:

- **Annotation language:** four sets of annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language was randomly selected for each of the images.
- **Caption fields:** only the fields for the *title*, *location*, *date* and additional *notes* were provided. Unlike 2006, the *description* field was not made available for retrieval to provide a more realistic evaluation scenario and to attract more visually oriented retrieval approaches.
- **Annotation completeness:** each image caption exhibited the same level of annotation completeness - there were no images without annotations as in 2006.

2.2 Query Topics

The participants were given 60 query topics (see Table 1) representing typical search requests for the generic photographic collection of the *IAPR TC-12 Benchmark*.

These topics had already been used in 2006, and we decided to reuse them to facilitate the objective comparison of retrieval from a generic collection of fully annotated (2006) and lightly annotated (2007) photographs. The creation of these topics is based on several factors (see [4] for detailed information), including:

- the analysis of a log file from online-access to the image collection;
- knowledge of the contents of the image collection;
- various types of linguistic and pictorial attributes;
- the use of geographical constraints;
- the estimated difficulty of the topic.

Similar to TREC, the query topics were provided as structured statements of user needs which consist of a title (a short sentence or phrase describing the search request in a few words) and three sample images that are relevant to that search request. These images were removed from the test collection and did not form part of the ground-truth in 2007.

The topic titles were offered in 16 languages including English, German, Spanish, Italian, French, Portuguese, Chinese, Japanese, Russian, Polish, Swedish, Finnish, Norwegian, Danish, and Dutch, whereby all translations had been provided by at least one native speaker and verified by at least another native speaker. The participants only received the topic titles, but not the narrative descriptions to avoid misunderstandings as they had been misinterpreted by participants in the past (they only serve to unambiguously define what constitutes a relevant image or not).

Table 1. *ImageCLEFphoto 2007* topics

ID Topic Title	ID Topic Title
1 accommodation with swimming pool	31 volcanos around Quito
2 church with more than two towers	32 photos of female guides
3 religious statue in the foreground	33 people on surfboards
4 group standing in front of mountain landscape in Patagonia	34 group pictures on a beach
5 animal swimming	35 bird flying
6 straight road in the USA	36 photos with Machu Picchu in the background
7 group standing in salt pan	37 sights along the Inca-Trail
8 host families posing for a photo	38 Machu Picchu and Huayna Picchu in bad weather
9 tourist accommodation near Lake Titicaca	39 people in bad weather
10 destinations in Venezuela	40 tourist destinations in bad weather
11 black and white photos of Russia	41 winter landscape in South America
12 people observing football match	42 pictures taken on Ayers Rock
13 exterior view of school building	43 sunset over water
14 scenes of footballers in action	44 mountains on mainland Australia
15 night shots of cathedrals	45 South American meat dishes
16 people in San Francisco	46 Asian women and/or girls
17 lighthouses at the sea	47 photos of heavy traffic in Asia
18 sport stadium outside Australia	48 vehicle in South Korea
19 exterior view of sport stadia	49 images of typical Australian animals
20 close-up photograph of an animal	50 indoor photos of churches or cathedrals
21 accommodation provided by host families	51 photos of goddaughters from Brazil
22 tennis player during rally	52 sports people with prizes
23 sport photos from California	53 views of walls with asymmetric stones
24 snowcapped buildings in Europe	54 famous television (and telecommunication) towers
25 people with a flag	55 drawings in Peruvian deserts
26 godson with baseball cap	56 photos of oxidised vehicles
27 motorcyclists racing at the Australian Motorcycle Grand Prix	57 photos of radio telescopes
28 cathedrals in Ecuador	58 seals near water
29 views of Sydney's world-famous landmarks	59 creative group pictures in Uyuni
30 room with more than two beds	60 salt heaps in salt pan

The participants were also given access to the results of a visual baseline run for each topic, provided by the FIRE system. The run thereby used colour histograms (compared with JSD, weight 3), Tamura texture histograms (compared with JSD, weight 2), and 32x32 thumbnails (compared with Euclidean distance, weight 1). More information on FIRE can be found in [5].

2.3 Relevance Assessments

Relevance assessments were carried out by the two topic creators using a custom-built online tool. The top 40 results from all submitted runs were used to create image pools giving an average of 2,299 images (max: 3237; min: 1513) 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. 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 sets of relevant images (qrels).

Finally, these qrels were complemented with the relevant images found at *ImageCLEFphoto 2006* in order to avoid missing out on relevant images not found this year due to the reduced captions.

2.4 Result Generation

Once the relevance judgments were completed, we were able to evaluate the performance of the individual systems and approaches. The results for submitted runs were computed using the latest version of *trec_eval*³ (Version 8.1).

The submissions were evaluated using uninterpolated (arithmetic) *mean average precisions* (MAP) and *precision at rank 20* (P20) because most online image retrieval engines like *Google*, *Yahoo!* and *Altavista* display 20 images by default. Further measures considered include *geometric mean average precision* (GMAP) to test system robustness, and the *binary preference* (bpref) measure which is an indicator for the completeness of relevance judgments.

3 Participation and Submission Overview

ImageCLEFphoto 2007 saw the registration of 32 groups (4 less than in 2006), with 20 of them eventually submitting 616 runs (all of which were evaluated). This is an increase in comparison to previous years (12 groups submitting 157 runs in 2006, and 11 groups submitting 349 runs in 2005 respectively).

Table 2. Participating groups

Group ID	Institution	Runs
Alicante	University of Alicante, Spain	6
Berkeley	University of California, Berkeley, USA	19
Budapest	Hungarian Academy of Sciences, Budapest, Hungary	11
CINDI	Concordia University, Montreal, Canada	5
CLAC	Concordia University, Montreal, Canada	6
CUT	Technical University Chemnitz, Germany	11
DCU-UTA	Dublin City University, Dublin, Ireland & University of Tampere, Finland	138
GE	University and Hospitals of Geneva, Switzerland	2
HongKong	Nanyang Technological University, Hong Kong	62
ImpColl	Imperial College, London, UK	5
INAOE	INAOE, Puebla, Mexico	115
IPAL	IPAL, Singapore	27
Miracle	Daedalus University, Madrid, Spain	153
NII	National Institute of Informatics, Tokyo, Japan	3
RUG	University of Groningen, The Netherlands	4
RWTH	RWTH Aachen University, Germany	10
SIG	Universite Paul Sabatier, Toulouse, France	9
SINAI	University of Jaén, Jaén, Spain	15
Taiwan	National Taiwan University, Taipei, Taiwan	27
XRCE	Cross-Content Analytics, Meylan, France	8

Table 2 provides an overview of the participating groups and the corresponding number of submitted runs. The 20 groups are from 16 countries, with one

³ http://trec.nist.gov/trec_eval/

institution (Concordia University) sending two separate groups (CINDI, CLAC), while DCU and UTA joined forces and submitted as one participating group. New participants submitting in 2007 include Budapest, CLAC, UTA, NTU (Hongkong), ImpColl, INAOE, RUG, SIG and XRCE. The number of runs per participating group has risen as well, with participants submitting an average of 30.8 runs in 2007 (13.1 runs in 2006). However, this may be attributed to the fact that four sets of annotations were offered (compared to two in 2007) and that the participants were allowed to submit as many runs as they desired.

The runs submitted were categorised with respect to the following dimensions: query and annotation language, run type (automatic or manual), use of relevance feedback or automatic query expansion, and modality (text only, image only or combined). Most submissions (91.6%) used the image annotations, with 8 groups submitting a total of 312 bilingual runs and 18 groups a total of 251 monolingual runs; 15 groups experimented with purely concept-based (textual) approaches (288 runs), 13 groups investigated the combination of content-based (visual) and concept-based features (276 runs), while a total of 12 groups submitted 52 purely content-based runs, an increase in comparison with previous events (in 2006, only 3 groups had submitted a total of 12 visual runs). Furthermore, 53.4% of all retrieval approaches involved the use of image retrieval (31% in 2006).

Based on all submitted runs, 50.6% were bilingual (59% in 2006), 54.7% of runs used query expansion and pseudo-relevance feedback techniques (or both) to further improve retrieval results (46% in 2006), and most runs were automatic (*i.e.* involving no human intervention); only 3.1% of the runs submitted were manual. Two participating groups made use of additional data (*i.e.* the description field and the qrels) from *ImageCLEFphoto 2006*. Although all these runs were evaluated (indicated by “Data 2006”), they were not considered for the system performance analysis and retrieval evaluation described in Section 4.

Table 3 displays the number of runs (and participating groups in parenthesis) with respect to query and annotation languages. The majority of runs (66.2%) was concerned with retrieval from English annotations, with exactly half of them (33.1%) being monolingual experiments and all groups (except for GE and RUG) submitting at least one monolingual English run. Participants also showed increased interest in retrieval from German annotations; a total of eight groups submitted 88 runs (14.5% of total runs), 20.5% of them monolingual (compared with four groups submitting 18 runs in 2006). Seven groups made use of the new Spanish annotations (5.4% of total runs, 48.5% of them monolingual), while only two participants experimented with the annotations with a randomly selected language for each image (5.3%).

The expanded multilingual character of the evaluation environment also yielded an increased number of bilingual retrieval experiments: while only four query languages (French, Italian, Japanese, Chinese) had been used in 10 or more bilingual runs in 2006, a total of 13 languages were used to start retrieval approaches in 10 or more runs in 2007. The most popular languages this year were German (43 runs), French (43 runs) and English (35 runs). Surprisingly, 26.5% of the bilingual experiments used a Scandinavian language to start the

Table 3. Submission overview by query and annotation languages

Query / Annotation	English	German	Spanish	Random	None	Total
English	204(18)	18 (5)	6 (3)	11 (2)		239(18)
German	31 (6)	31 (7)	1 (1)	11 (2)		74 (9)
Visual	1 (1)				52 (12)	53(12)
French	32 (7)	1 (1)	10 (2)			43 (7)
Spanish	20 (5)		16 (7)	2 (1)		38 (9)
Swedish	20 (3)	12 (1)				32 (3)
Simplified Chinese	24 (4)	1 (1)				25 (4)
Portuguese	19 (5)			2 (1)		21 (5)
Russian	17 (4)	1 (1)		2 (1)		20 (4)
Norwegian	6 (1)	12 (1)				18 (1)
Japanese	16 (3)					16 (3)
Italian	10 (4)			2 (1)		12 (4)
Danish		12 (1)				12 (1)
Dutch	4 (1)			2 (1)		6 (1)
Traditional Chinese	4 (1)					4 (1)
Total	408 (18)	88 (8)	33 (7)	32 (2)	52 (12)	616(20)

retrieval approach: Swedish (32 runs), Norwegian (18 runs) and Danish (12 runs) – none of these languages had been used in 2006. It is also interesting to note that Asian languages (18.6% of bilingual runs) were almost exclusively used for retrieval from English annotations (only one run experimented with the German annotations), which might indicate a lack of translation resources from Asian to European languages other than English.

4 Results

This section provides an overview of the system results with respect to query and annotation languages as well as other submission dimensions such as query mode, retrieval modality and the involvement of relevance feedback or query expansion techniques. Although the description fields were not provided with the image annotations, the absolute retrieval results achieved by the systems were not much lower compared to those in 2006 when the entire annotation was used. We attribute this to the fact that more than 50% of the groups had participated at ImageCLEF before, improved retrieval algorithms (not only of returning participants), and the increased use of content-based retrieval approaches.

4.1 Results by Language

Table 4 shows the runs which achieved the highest MAP for each language pair (ranked by descending order of MAP scores).

Of these runs, 90.6% use query expansion or relevance feedback, and 78.1% use both visual and textual features for retrieval. It is noticeable that submissions from CUT, DCU, NTU (Taiwan) and INAOE dominate the results. As

Table 4. Systems with highest MAP for each language

Query (Caption)	Group/Run ID	MAP	P(20)	GMAP	bpref
English (English)	CUT/cut-EN2EN-F50	0.318	0.459	0.298	0.162
German (English)	XRCE/DE-EN-AUTO-FB-TXTIMG_MPRF	0.290	0.388	0.268	0.156
Portuguese (English)	Taiwan/NTU-PT-EN-AUTO-FBQE-TXTIMG	0.282	0.388	0.266	0.127
Spanish (English)	Taiwan/NTU-ES-EN-AUTO-FBQE-TXTIMG	0.279	0.383	0.259	0.128
Russian (English)	Taiwan/NTU-RU-EN-AUTO-FBQE-TXTIMG	0.273	0.383	0.256	0.115
Italian (English)	Taiwan/NTU-IT-EN-AUTO-FBQE-TXTIMG	0.271	0.384	0.257	0.114
S. Chinese (English)	CUT/cut-ZHS2EN-F20	0.269	0.404	0.244	0.098
French (English)	Taiwan/NTU-FR-EN-AUTO-FBQE-TXTIMG	0.267	0.374	0.248	0.115
T. Chinese (English)	Taiwan/NTU-ZHT-EN-AUTO-FBQE-TXTIMG	0.257	0.360	0.240	0.089
Japanese (English)	Taiwan/NTU-JA-EN-AUTO-FBQE-TXTIMG	0.255	0.368	0.241	0.094
Dutch (English)	INAOE/INAOE-NL-EN-NaiveWBQE-IMFB	0.199	0.292	0.191	0.038
Swedish (English)	INAOE/INAOE-SV-EN-NaiveWBQE-IMFB	0.199	0.292	0.191	0.038
Visual (English)	INAOE/INAOE-VISUAL-EN-AN_EXP_3	0.193	0.294	0.192	0.039
Norwegian (English)	DCU/NO-EN-Mix-sgramRF-dyn-equal-fire	0.165	0.275	0.174	0.057
German (German)	Taiwan/NTU-DE-DE-AUTO-FBQE-TXTIMG	0.245	0.379	0.239	0.108
English (German)	XRCE/EN-DE-AUTO-FB-TXTIMG_MPRF_FLR	0.278	0.362	0.250	0.112
Swedish (German)	DCU/SW-DE-Mix-dictRF-dyn-equal-fire	0.179	0.294	0.180	0.071
Danish (German)	DCU/DA-DE-Mix-dictRF-dyn-equal-fire	0.173	0.294	0.176	0.073
French (German)	CUT/cut-FR2DE-F20	0.164	0.237	0.144	0.004
Norwegian (German)	DCU/NO-DE-Mix-dictRF-dyn-equal-fire	0.167	0.270	0.165	0.070
Spanish (Spanish)	Taiwan/NTU-ES-ES-AUTO-FBQE-TXTIMG	0.279	0.397	0.269	0.113
English (Spanish)	CUT/cut-EN2ES-F20	0.277	0.377	0.247	0.105
German (Spanish)	Berkeley/Berk-DE-ES-AUTO-FB-TXT	0.091	0.122	0.072	0.008
English (Random)	DCU/EN-RND-Mix-sgramRF-dyn-equal-fire	0.168	0.285	0.175	0.068
German (Random)	DCU/DE-RND-Mix-sgram-dyn-equal-fire	0.157	0.282	0.167	0.064
French (Random)	DCU/FR-RND-Mix-sgram-dyn-equal-fire	0.141	0.264	0.148	0.059
Spanish (Random)	INAOE/INAOE-ES-RND-NaiveQE-IMFB	0.124	0.228	0.136	0.027
Dutch (Random)	INAOE/INAOE-NL-RND-NaiveQE	0.083	0.156	0.094	0.011
Italian (Random)	INAOE/INAOE-IT-RND-NaiveQE	0.080	0.144	0.086	0.018
Russian (Random)	INAOE/INAOE-RU-RND-NaiveQE	0.076	0.136	0.085	0.017
Portuguese (Random)	INAOE/INAOE-PT-RND-NaiveQE	0.030	0.043	0.032	0.001
Visual	XRCE/AUTO-NOFB-IMG_COMBFBK	0.189	0.352	0.201	0.102

in previous years, the highest English monolingual run outperforms the highest German and Spanish monolingual runs (MAPs are 22.9% and 12.1% lower).

The highest bilingual to English run (German – English) performed with a MAP of 91.3% of the highest monolingual run MAP, with the highest bilingual run in most other query languages such as Portuguese, Spanish, Russian, Italian, Chinese, French and Japanese all exhibiting at least 80% of that highest monolingual English run. Hence, there is no longer much difference between monolingual and bilingual retrieval, indicating a significant progress of the translation and retrieval methods using these languages. Moreover, the highest bilingual to Spanish run (English – Spanish) had a MAP of 99.2% of the highest monolingual Spanish run, while the highest bilingual to German run (English – German) even outperformed the highest German monolingual run MAP by 13.3%.

4.2 Results by Query Mode

This trend is not only true for the highest runs per language pair, but also for all submissions and across several performance indicators. Table 5 illustrates the average scores across all system runs (and the standard deviations in parenthesis) with respect to monolingual, bilingual and purely visual retrieval.

Again, monolingual and bilingual retrieval are almost identical, and so are the average results for monolingual Spanish, English and German retrieval (see

Table 5. Results by query mode

Query Mode	MAP	P(20)	BPREF	GMAP
Monolingual	0.138 (0.070)	0.192 (0.102)	0.132 (0.066)	0.038 (0.036)
Bilingual	0.136 (0.056)	0.199 (0.088)	0.136 (0.054)	0.037 (0.027)
Visual	0.068 (0.039)	0.157 (0.069)	0.080 (0.039)	0.022 (0.019)

Table 6): Spanish shows the highest average MAP and BPREF values, while German exhibits the highest average for P(20) and English for GMAP.

Table 6. Monolingual results by annotation language

Annotation	MAP	P(20)	BPREF	GMAP
Spanish	0.145 (0.059)	0.195 (0.092)	0.134 (0.056)	0.036 (0.034)
English	0.139 (0.075)	0.190 (0.108)	0.132 (0.071)	0.038 (0.038)
German	0.133 (0.043)	0.200 (0.083)	0.132 (0.048)	0.034 (0.031)

Across all submissions, the average values for bilingual retrieval from English and German annotations are even slightly higher than those for monolingual retrieval (see Table 7), while bilingual retrieval from Spanish annotations and from annotations with a randomly selected language does not lag far behind.

Table 7. Bilingual results by annotation language

Annotation	MAP	P(20)	BPREF	GMAP
English	0.150 (0.055)	0.204 (0.089)	0.143 (0.054)	0.037 (0.029)
German	0.138 (0.040)	0.217 (0.075)	0.145 (0.040)	0.045 (0.021)
Spanish	0.117 (0.079)	0.176 (0.108)	0.108 (0.070)	0.027 (0.037)
Random	0.099 (0.048)	0.169 (0.084)	0.108 (0.052)	0.028 (0.021)
None	0.068 (0.039)	0.157 (0.069)	0.080 (0.039)	0.022 (0.019)

These results indicate that the query language does not play a major factor for visual information retrieval for lightly annotated images. We attribute this (1) to the high quality of the state-of-the-art translation techniques, (2) to the fact that such translations implicitly expand the query terms (similar to query expansion using a thesaurus) and (3) to the short image captions used (as many of them are proper nouns which are often not even translated).

4.3 Results by Retrieval Modality

In 2006, the system results had shown that combining visual features from the image and semantic knowledge derived from the captions offered optimum performance for retrieval from a generic photographic collection with fully annotated images.

Table 8. Results by retrieval modality

Modality	MAP	P(20)	BPREF	GMAP
Mixed	0.149 (0.066)	0.225 (0.097)	0.203 (0.081)	0.050 (0.031)
Text Only	0.120 (0.040)	0.152 (0.051)	0.141 (0.045)	0.018 (0.018)
Image Only	0.068 (0.039)	0.157 (0.069)	0.080 (0.039)	0.022 (0.019)

As indicated in Table 8, the results of *ImageCLEFphoto 2007* show that this also applies for retrieval from generic photographic collections with lightly annotated images: on average, combining visual features from the image and semantic information from the annotations gave a 24% improvement of the MAP over retrieval based solely on text.

Purely content-based approaches still lag behind, but the average MAP for retrieval solely based on image features shows an improvement of 65.8% compared to the average MAP in 2006.

4.4 Results by Feedback and/or Query Expansion

Table 9 illustrates the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the use of query expansion or relevance feedback techniques.

Table 9. Results by feedback or query expansion

Technique	MAP	P(20)	BPREF	GMAP
None	0.109 (0.052)	0.178 (0.075)	0.110 (0.047)	0.027 (0.024)
Query Expansion	0.112 (0.040)	0.158 (0.053)	0.106 (0.036)	0.024 (0.019)
Relevance Feedback	0.131 (0.055)	0.185 (0.084)	0.132 (0.054)	0.038 (0.026)
Expansion & Feedback	0.218 (0.062)	0.324 (0.076)	0.209 (0.053)	0.073 (0.046)

While the use of query expansion (*i.e.* the use of thesauri or ontologies such as WordNet) does not necessarily seem to dramatically improve retrieval results for retrieval from lightly annotated images (average MAP only 2.1% higher), relevance feedback (typically in the form of query expansion based on pseudo relevance feedback) appeared to work well on short captions (average MAP 19.9% higher), with a combination of query expansion and relevance feedback techniques yielding results almost twice as good as without any of these techniques (average MAP 99.5% higher).

4.5 Results by Run Type

Table 10 shows the average scores across all systems runs (and the standard deviations in parenthesis) with respect to the run type. Unsurprisingly, MAP results of manual approaches are, on average, 58.6% higher than purely automatic runs — this trend seems to be true for both fully annotated and lightly annotated images.

Table 10. Results by run type

Technique	MAP	P(20)	BPREF	GMAP
Manual	0.201 (0.081)	0.302 (0.116)	0.189 (0.074)	0.066 (0.051)
Automatic	0.127 (0.058)	0.187 (0.084)	0.126 (0.055)	0.034 (0.029)

5 Conclusion

This paper reported on *ImageCLEFphoto 2007*, the general photographic ad-hoc retrieval task of the *ImageCLEF 2007* evaluation campaign. Its evaluation objective concentrated on visual information retrieval from generic collections of lightly annotated images, a new challenge that attracted a large number of submissions: 20 participating groups submitted a total of 616 system runs.

The participants were provided with a subset of the *IAPR TC-12 Benchmark*: 20,000 colour photographs and four sets of semi-structured annotations in (1) English, (2) German, (3) Spanish and (4) one set whereby the annotation language was randomly selected for each of the images. Unlike in 2006, the participants were not allowed to use the semantic description field in their retrieval approaches. The topics and relevance assessments from 2006 were reused (and updated) to facilitate the comparison of retrieval from fully and lightly annotated images.

The nature of the task also attracted a larger number of participants experimenting with content-based retrieval techniques, and hence the retrieval results were similar to those in 2006, despite the limited image annotations in 2007. Other findings for multilingual visual information retrieval from generic collections of lightly annotated photographs include:

- bilingual retrieval performs as well as monolingual retrieval;
- the choice of the query language is almost negligible as many of the short captions contain proper nouns;
- combining concept and content-based retrieval methods as well as using relevance feedback and/or query expansion techniques can significantly improve retrieval performance;

ImageCLEFphoto will continue to provide resources to the retrieval and computational vision communities to facilitate standardised laboratory-style testing of image retrieval systems. While these resources have predominately been used by systems applying a concept-based retrieval approach thus far, the rapid increase of participants using content-based retrieval techniques at *ImageCLEFphoto* calls for a more suitable evaluation environment for visual approaches (*e.g.* the preparation of training data). For *ImageCLEFphoto 2008*, we are planning to create new topics and will therefore be able to provide this year's topics and qrels as training data for next year.

Acknowledgements

We would like to thank *viventura* and the *IAPR TC-12* for providing their image databases for this year's task. This work was partially funded by the EU MultiMatch project (IST-033104) and the EU MUSCLE NoE (FP6-507752).

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.D., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 579–594. Springer, Heidelberg (2007)
3. Grubinger, M., Clough, P.D., Müller, H., Deselaers, T.: The IAPRTC12 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 2006, Genoa, Italy, pp. 13–23 (2006)
4. Grubinger, M.: On the Creation of Query Topics for ImageCLEFphoto. In: Third MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation, Budapest, Hungary, pp. 50–63 (2007)
5. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: An experimental comparison. Information Retrieval (in press, 2008)

Overview of the ImageCLEF 2007 Object Retrieval Task

Thomas Deselaers¹, Allan Hanbury², Ville Viitaniemi³, András Benczúr⁴,
Mátyás Brendel⁴, Bálint Daróczy⁴, Hugo Jair Escalante Balderas⁵,
Theo Gevers⁶, Carlos Arturo Hernández Gracidas⁵, Steven C.H. Hoi⁷,
Jorma Laaksonen³, Mingjing Li⁸, Heidy Marisol Marín Castro⁵,
Hermann Ney¹, Xiaoguang Rui⁸, Nicu Sebe⁶, Julian Stöttinger², and Lei Wu⁸

¹ Computer Science Department, RWTH Aachen University, Germany
`deselaers@cs.rwth-aachen.de`

² Pattern Recognition and Image Processing Group (PRIP), Institute of
Computer-Aided Automation, Vienna University of Technology, Austria

³ Adaptive Informatics Research Centre, Helsinki University of Technology, Finland

⁴ Data Mining and Web search Research Group, Computer and Automation Research
Institute of the Hungarian Academy of Sciences, Budapest, Hungary

⁵ TIA Research Group, Computer Science Department, National Institute of
Astrophysics, Optics and Electronics, Tonantzintla, Mexico

⁶ Intelligent Systems Lab Amsterdam, University of Amsterdam, The Netherlands

⁷ School of Computer Engineering, Nanyang Technological University, Singapore

⁸ Microsoft Research Asia, Beijing, China

Abstract. We describe the object retrieval task of ImageCLEF 2007, give an overview of the methods of the participating groups, and present and discuss the results.

The task was based on the widely used *PASCAL object recognition data* to train object recognition methods and on the *IAPR TC-12 benchmark dataset* from which images of objects of the ten different classes bicycles, buses, cars, motorbikes, cats, cows, dogs, horses, sheep, and persons had to be retrieved.

Seven international groups participated using a wide variety of methods. The results of the evaluation show that the task was very challenging and that different methods for relevance assessment can have a strong influence on the results of an evaluation.

1 Introduction

Object class recognition, automatic image annotation, and object retrieval are strongly related tasks. In object class recognition, the aim is to identify whether a certain object is contained in an image; in automatic image annotation, the aim is to create a textual description of a given image; and in object retrieval, images containing certain objects or object classes have to be retrieved out of a large set of images. Each of these techniques is important to allow for semantic retrieval from image collections.

Over the last year, research in these areas has strongly grown, and it is becoming clear that performance evaluation is a very important component for fostering progress in research. Several initiatives create benchmark suites and databases to quantitatively compare different methods tackling the same problem.

In the last years, evaluation campaigns for object detection [1,2], content-based image retrieval [3] and image classification [4] have developed. There is however, no task aiming at finding images showing a particular object from a larger database. Although this task is extremely similar to the PASCAL visual object classes challenge [1,2], it is not the same. In the PASCAL object recognition challenge, the probability for an object to be contained in an image is relatively high and the images to train and test the methods are from the same data collection. In realistic scenarios, this might not be a suitable assumption. Therefore, in the object retrieval task described here, we use the training data that was carefully assembled by the PASCAL NoE with much manual work, and the IAPR TC-12 database which has been created under completely different circumstances as the database from which relevant images are to be retrieved.

In this paper, we present the results of the object retrieval task that was arranged as part of the CLEF/ImageCLEF 2007 image retrieval evaluation. This task was conceived as a purely visual task, making it inherently cross-lingual. Once one has a model for the visual appearance of a specific object, such as a bicycle, it can be used to find images of bicycles independently of the language or quality of the annotation of an image.

ImageCLEF¹ [3] started within CLEF² (Cross Language Evaluation Forum) in 2003. A medical image retrieval task was added in 2004 to explore domain-specific multilingual information retrieval and also multi-modal retrieval by combining visual and textual features for retrieval. Since 2005, a medical retrieval and a medical image annotation task are both part of ImageCLEF. In 2006, a general object recognition task was presented to see whether interest in this area existed. Although only a few groups participated, many groups expressed their interest and encouraged us to create an object retrieval task. In ImageCLEF 2007, aside from the object retrieval task described here, a photographic retrieval task also using the IAPR TC-12 database [5], a medical image retrieval task [6], and a medical automatic annotation task [6] were organised.

2 Task Description

The task was defined as a visual object retrieval task. Training data was in the form of annotated example images of ten object classes (PASCAL VOC 2006 data). The task was to learn from the provided annotated images and then to find all images in the IAPR-TC12 database containing the learned objects. The particularity of the task is that the training and test images are not from the same set of images. This makes the task more realistic, but also more challenging.

¹ <http://www.imageclef.org>

² <http://www.clef-campaign.org/>

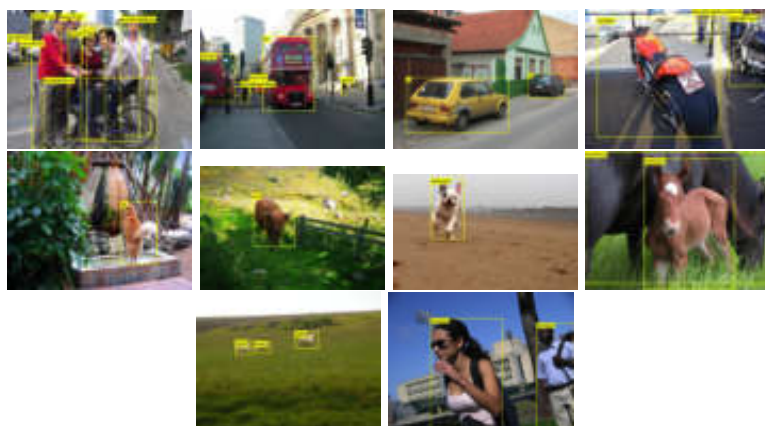


Fig. 1. Example images from the PASCAL VOC 2006 training dataset

2.1 Datasets

For this task, the two datasets described below were used:

PASCAL VOC 2006 training data: As training data, the organisers of the PASCAL Network of Excellence visual object classes (VOC) challenge kindly agreed that we use the training data they assembled for their 2006 challenge. This data is freely available on the PASCAL web-page³ and consists of approximately 2600 images, where for each image a detailed description of which of the ten object classes is visible in which area of the image is available (indicated by bounding boxes). Example images from this database are shown in Figure 1 with the corresponding annotation.

IAPR TC-12 dataset: The IAPR TC-12 Benchmark database [7] consists of 20,000 still images taken from locations around the world and comprising an assorted cross-section of still images which might for example be found in a personal photo collection. It includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Some example images are shown in Figure 2. This data is also strongly annotated using textual descriptions of the images and various meta-data. We use only the image data for this task.

2.2 Object Retrieval Task

The ten queries correspond to the ten classes of the PASCAL VOC 2006 data: bicycles, buses, cars, motorbikes, cats, cows, dogs, horses, sheep, and persons. For training, only the “train” and “val” sections of the PASCAL VOC database were to be used. For each query, participants were asked to submit a list of 1000 images obtained by their method from the IAPR-TC12 database, ranked in the order of best to worst satisfaction of the query.

³ <http://www.pascal-network.org/challenges/VOC/>

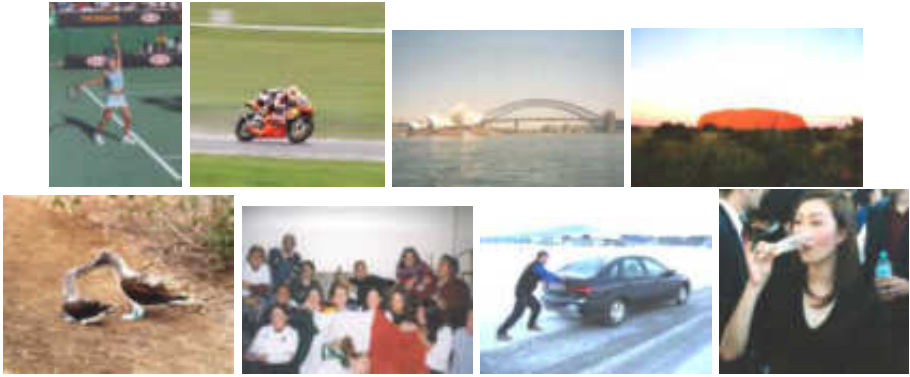


Fig. 2. Example images from the IAPR TC-12 benchmark dataset

2.3 Evaluation Measure

To evaluate the retrieval performance we use the same measure used by most retrieval evaluations such as the other tasks in CLEF/ImageCLEF [5,6], TREC⁴ and TRECVID⁵. The *average precision (AP)* gives an indication of the retrieval quality for one topic and the *mean average precision (MAP)* provides a single-figure measure of quality across recall levels averaged over all queries. To calculate these measures, it of course necessary to judge which images are relevant for a given query and which are not. To calculate the evaluation measures we use `trec_eval`⁶, the standard program from TREC.

2.4 Relevance Assessments

To find relevant images, we created pools per topic [8] keeping the top 100 results from all submitted runs resulting in 1,507 images to be judged per topic on average. This resulted in a total of 15,007 images to be assessed. The normal relevance judgement process in information retrieval tasks envisages that several users judge each document in question for relevance and that for each image relevance for the particular query is judged. Given that judging the presence or absence of a given object in an image is a straightforward task, we postulate that every two persons among the judges would come to the same conclusion, and therefore each image was judged by only one judge. The whole judgement process was performed over a web interface which was quickly created and everybody from the RWTH Aachen University Human Language Technology and from the Vienna University of Technology Pattern Recognition and Image Processing (PRIP) group was invited to judge images. Thus, most of the judges are computer science students and researchers with a human language technology or pattern Recognition and

⁴ <http://trec.nist.gov/>

⁵ <http://www-nlpir.nist.gov/projects/t01v/>

⁶ http://trec.nist.gov/trec_eval/

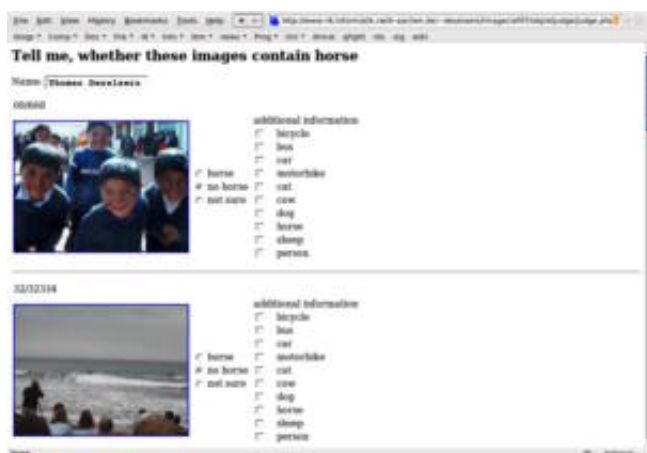


Fig. 3. The relevance judgement web-interface

image analysis background. Note that in the pooling process all images that are not judged are automatically considered to be not relevant.

The web-interface is shown in Figure 3 to give an impression of the process. On each page, 10 images are shown, and the judge has to decide whether a particular object is present in these images or not. To reduce boredom for the judges, they are allowed (and recommended) to specify whether other object classes are present in the images. This facility was added after the first 3,000 images had already been judged due to complaints by the judges that the task was too simple. The judges were told to be rather positive about the relevance of an image, e.g. to consider sheep-like animals such as llamas to be sheep and to consider tigers and other non-domestic cats to be cats. In the analysis of the results published in [9] and [10] it turned out that these judging guidelines were rather to imprecise and led to an inconsistent judging of the images.

Furthermore, Ville Viitaniemi from the HUTCIS group, judged all 20,000 images with respect to relevance for all of the topics with a stricter definition of relevances.

Results from the Relevance Judgements. Table 1 gives an overview how many images were found to be relevant for each of the given topics using simulated pooling. For the initial evaluation [9], the pooling was done using some incorrect submissions and without sufficiently strict judging guidelines. Here, the pooling was simulated after all runs were checked to strictly follow the submission guidelines using the annotation of the full database. It can be observed that there are far more relevant images for the person topic than for any other topic. From these numbers it can be seen that the task at hand is challenging for most of the classes. It can also be observed that the percentage of relevant images in the additional pooling observation is very similar to the full database annotation and thus we can assume that choosing a (sufficiently large) random partition

Table 1. Results from the relevance judgement process. Column 3 shows the number of relevant images when standard (simulated) pooling is used, column 4 when the (simulated) additional class information is taken into account. Column 5 shows the results of the relevance judgement of all 20,000 images.

query	query name	relev. in pool	additional relev.	relev. in database
1	bicycle	81/1422 (5.7%)	350/10060 (3.5%)	655/20000 (3.3%)
2	bus	29/1481 (2.0%)	106/10060 (1.1%)	218/20000 (1.1%)
3	car	219/1665 (13%)	644/10060 (6.4%)	1268/20000 (6.3%)
4	motorbike	17/1481 (1.1%)	48/10060 (0.48%)	86/20000 (0.43%)
5	cat	2/1566 (0.13%)	4/10060 (0.04%)	7/20000 (0.04%)
6	cow	10/1559 (0.64%)	30/10060 (0.30%)	49/20000 (0.25%)
7	dog	4/1554 (0.26%)	32/10060 (0.32%)	72/20000 (0.36%)
8	horse	33/1547 (2.1%)	110/10060 (1.1%)	175/20000 (0.88%)
9	sheep	0/1427 (0.00%)	1/10060 (0.01%)	6/20000 (0.03%)
10	person	1095/1734 (63%)	5356/10060 (53%)	11248/20000 (56%)

of documents to be judged can lead to a good estimate of relevant documents in the database. However, since the assumption that objects occur uncorrelated in the images is certainly invalid, this additional relevance information, which favors images with at least two different objects shown, is not optimal.

If only the data from the conventional pooling process is considered, then for five of the ten classes less than a thousandth of all images in the database are relevant, and the fact that still a high number of images has to be judged makes the usefulness of the whole judging process for this task questionable.

Another problem with pooling is reusability: since only a small portion of the relevant images in the whole database is found by the pooling process, the evaluation of a new method with the found pools is questionable. The additional pools, given that more of the relevant images are found, might be better suited, but as described above introduce a different form of bias.

3 Methods

Seven international groups from academia participated in the task and submitted a total of 38 runs. The group with the highest number of submissions had 13 submissions. In the following sections, the methods of the groups are explained (in alphabetical order) and references to further work are given.

3.1 Budapest Methods

Authors: Mátyás Brendel, Bálint Daróczy, and András Benczúr

Affiliation: Data Mining and Web search Research Group, Informatics Laboratory, Computer and Automation Research Institute of the Hungarian Academy of Sciences

Email: {mbrendel,daroczyb,benczur}@ilab.sztaki.hu

Budapest-Acad315. The task of object retrieval is to classify objects found in images. This means to find objects in an image that are similar to sample objects in the pre-classified images. There are two problems with this task: the first is, how do we model objects. The second is, how do we measure similarity of objects. Our first answer to the first question is to model objects with image segments. Segment, region or blob based image similarity is a common method in content based image retrieval, see for example [11,12,13,14].

Instead of the PASCAL VOC 2006 database we used the PASCAL VOC 2007 database, since that database contained samples with exact object-boundaries, which is important for our methods. It is possible that our method will also work almost with the same efficiency with the PASCAL VOC 2006 database, but we have no test for this at current time.

The basis of our first method is to find segments on the query image which are similar to the objects in the pre-classified images. The image is then classified to be in that class, to which we find the most similar segment in the query image.

Image segmentation in itself is a widely researched and open problem. We used an image segmenter developed by our group to extract segments from the query images. Our method is based on a graph-based algorithm developed by Felzenszwalb and Huttenlocher [15]. We implemented a pre-segmentation method to reduce the computational time and use a different smoothing technique. All images were sized to a fixed resolution. Gaussian-based smoothing helped us cut down high frequency noise. Because of the efficiency of the OpenCV⁷, implementation we did not implement resizing and Gaussian-based smoothing algorithms. As pre-segmentation we built a three-level Gaussian-Laplacian pyramid to define initial pixel groups. The original pyramid-based method, which considers the connection between pixels on different levels too, was modified to eliminate the so-called blocking problem. We used brightness difference to measure distance between pixels:

$$diffY(P_1, P_2) = 0.3 * |R_{P_2} - R_{P_1}| + 0.59 * |G_{P_2} - G_{P_1}| + 0.11 * |B_{P_2} - B_{P_1}| \quad (1)$$

After pre-segmentation, we had segments of 16 pixels maximum. To detect complex segments, we modified the original graph-based method by Felzenszwalb and Huttenlocher [15] with an adaptive threshold system using Euclidean distance to prefer larger regions instead of small regions of the image. Felzenszwalb and Huttenlocher defined an undirected graph $G = (V, E)$ where $\forall v_i \in V$ corresponds to a pixel in the image, and the edges in E connect certain pairs of neighboring pixels. This graph-based representation of the image reduces the original proposition into a graph cutting challenge. They made a very efficient and linear algorithm that yields a result near to the optimal normalized cut which is one of the NP-complete graph problems [15,16]. The algorithm is listed in Algorithm 1.

This algorithm sometimes does not find relevant parts with low initial thresholds. To find the relevant borders which would disappear with the graph-based

⁷ <http://www.intel.com/technology/computing/opencv/>

Algorithm 1. Segmentation algorithm.

Algorithm *Segmentation* (I_{src}, τ_1, τ_2)
 τ_1 and τ_2 are threshold functions. Let I_2 be the source image, I_1 and I_0 are the down-scaled images. Let $P(x, y, i)$ be the pixel $P(x, y)$ in the image on level i (I_i). Let $G = (V, E)$ be an undirected weighted graph where $\forall v_i \in V$ corresponds to a pixel $P(x, y)$. Each edge (v_i, v_j) has a non-negative weight $w(v_i, v_j)$.

Gaussian-Laplacian Pyramid

1. For every $P(x, y, 1)$ $Join(P(x, y, 1), P(x/2, y/2, 0))$ if $\tau_1 < diffY(P(x, y, 1), P(x/2, y/2, 0))$
2. For every $P(x, y, 2)$ $Join(P(x, y, 2), P(x/2, y/2, 1))$ if $\tau_1 < diffY(P(x, y, 2), P(x/2, y/2, 1))$

Graph-based Segmentation

1. Compute $Max_{weight}(R) = \max_{e \in MST(R, E)} w(e)$ for every coherent group of points R where $MST(R, E)$ is the minimal spanning tree
 2. Compute $Co(R) = \tau_2(R) + Max_{weight}(R)$ as the measure of coherence between points in R
 3. $Join(R_1, R_2)$ if $e \in E$ exists so $w(e) < \min(Co(R_1), Co(R_2))$ is true, where $R_1 \cap R_2 = \emptyset$ and $w(e)$ is the weight of the border edge e between R_1 and R_2
 4. Repeat steps 1,2,3 for every neighboring group (R_1, R_2) until possible to join two groups
-

method using high thresholds we calculated the Sobel-gradient image to separate important edges from other remainders.

Similarity of complex objects is usually measured on a feature base. This means that the similarity of the objects is defined by the similarity in a certain feature space.

$$dist(S_i, O_j) = d(F(S_i), F(O_j)) : S_i \in S, O_j \in O \quad (2)$$

where S is the set of segments and O is the set of objects, $dist$ is the distance function of the objects and segments, d is a distance function in the feature space (usually some of the conventional metrics in the n -dimensional real space), F is the function which assigns features to objects and segments. We extracted from the segments features, like mean color, size, shape information, and histogram information. As shape information a 4×4 sized low-resolution variant of the segment (framed in a rectangle with background) was used. Our histograms had 5 bins in each channel. Altogether a 35 dimensional, real valued feature-vector was extracted for each of the segments. The same features were extracted for the objects in the pre-classified images taking them as segments. The background and those classes which were not requested were ignored. The features of the objects were written to a file, with the class-identifiers, which were extracted from the color-coding. This way we obtained a data-base of class samples, containing features of objects belonging to the classes. After this, the comparison of the objects of the pre-classified sample images and the segments of the query image was possible. We used Euclidean distance to measure similarity. The distance of the query-image Q was computed as:

$$dist(Q) = \min_{i,j} dist(S_i, O_j) : S_i \in S, O_j \in O \quad (3)$$

where S is the set of segments of image Q , O is the set of the pre-classified sample objects. Q is classified to be in the class of the object that minimizes the distance. The score of an image was computed as:

$$score(Q) = 1000/dist(Q) \quad (4)$$

where Q is the query image.

Budapest-Acad314. In our first method (see budapest-acad315) we found that our segments are much smaller than the objects in the pre-segmented images. It would have been possible to get larger segments by adjusting the segmentation algorithm, however this way we would not get segments which were really similar to the objects. We found that our segmentation algorithm could not generate segments similar to the the objects in the pre-classified images with any settings of the parameters. Even if we tried our algorithm on the sample images, and the segments were approximately of the same size, the segments did not match the pre-classified objects. The reason for this is that pre-segmentation was made by humans and algorithmic segmentation is far from capable of the same result. For example, it is almost impossible to write an algorithm, which would segment a shape of a human being as one segment if his clothes are different. However, people were one of the classes defined, and the sample images contained people with the entire body as one object. Therefore we modified our method. Our second method is still segment-based. But we also do a segmentation on the sample-images. We took the segmented sample-images, and if a segment was 80% inside of an area of a pre-defined object, then we took this segment as a proper sample for that object. This way a set of sample segments was created. After this the method is similar to the previous, the difference is only that we have sample segments instead of sample objects, but we treat them the same way. The features of the segments were extracted and they were written to a file, with the identifier of the class, which was extracted from the color-codes. After this, the comparison of the segments of the pre-classified images and the query image was possible. We used Euclidean distance again to measure similarity. The closest segment of the image to a segment in any of the objects was searched using thhe distance

$$dist(Q) = \min_{i,j} dist(S_i, S_j) : S_i \in S, S_j \in O \quad (5)$$

where S is the segments of image Q , O is the set of segments belonging to the pre-classified objects. The image was classified according to the object, to which the closest segment belongs. As we expected, this modification made the algorithm better.

3.2 HUTCIS: Conventional Supervised Learning Using Fusion of Image Features

Authors: Ville Viitaniemi, Jorma Laaksonen

Affiliation: Adaptive Informatics Research Centre/Laboratory of Computer and Information Science, Helsinki University of Technology, Finland

Email: `firstname.lastname@tkk.fi`

All our 13 runs identified with prefix HUTCIS implement a similar general system architecture with three system stages:

1. Extraction of a large number of global and semi-global image features. Here we interpret global histograms of local descriptors as one type of global image feature.
2. For each individual feature, conventional supervised classification of the test images using the VOC2006 trainval images as the training set.
3. Fusion of the feature-wise classifier outputs.

By using this architecture, we knowingly ignored the aspect of qualitatively different training and test data. The motivation was to provide a baseline performance level that could be achieved by just applying a well-working implementation of the conventional supervised learning approach. Table 2 with ROC AUC performances in the VOC 2006 test set reveals that the performance of our principal run HUTCIS_SVM_FULLIMG_ALL is relatively close to the best performances in last year's VOC evaluation [2]. The last row of the table indicates what the rank of the HUTCIS_SVM_FULLIMG_ALL run would have been among the 19 VOC 2006 participants.

The following briefly describes the components of the architecture. For a more detailed description, see e.g. [17].

Features: For different runs, the features are chosen from a set of feature vectors, each with several components. Table 3 lists 10 of the features. Additionally, the available feature set includes interest point SIFT feature histograms with different histogram sizes, and concatenations of pairs, triples and quadruples of the tabulated basic feature vectors. The SIFT histogram bins have been selected by clustering part of the images with the self-organising map (SOM) algorithm.

Classification and fusion: The classification is performed either by a C-SVC implementation built around the LIBSVM support vector machine (SVM) library [18], or a SOM-based classifier [19]. The SVM classifiers (prefix HUTCIS_SVM) are fused together using an additional SVM layer. For the SOM classifiers (prefix HUTCIS_PICSOM), the fusion is based on the summation of the normalised classifier outputs.

The different runs: Our principal run HUTCIS_SVM_FULLIMG_ALL implements all the three system stages in the best way possible. Other runs use

Table 2. ROC AUC performance in VOC2006 test set

Run id.	bic.	bus	car	cat	cow	dog	horse	mbike	person	sheep
FULLIMG_ALL	0.921	0.978	0.974	0.930	0.937	0.866	0.932	0.958	0.874	0.941
FULLIMG_IP+SC	0.922	0.977	0.974	0.924	0.934	0.851	0.928	0.953	0.865	0.941
FULLIMG_IP	0.919	0.952	0.970	0.917	0.926	0.840	0.903	0.943	0.834	0.936
Best in VOC2006	0.948	0.984	0.977	0.937	0.940	0.876	0.927	0.969	0.863	0.956
Rank	7th	4th	3rd	4th	4th	3rd	1st	5th	1st	6th

Table 3. Some of the image features used in the HUTCIS runs

Colour layout	Dominant colour
Sobel edge histogram (4x4 tiling of the image)	HSV colour histogram
Average colour (5-part tiling)	Colour moments (5-part tiling)
16 × 16 FFT of edge image	Sobel edge histogram (5-part tiling)
Sobel edge co-occurrence matrix (5-part tiling)	

subsets of the image features, inferior algorithms or are otherwise predicted to be suboptimal.

The run HUTCIS_SVM_FULLIMG_ALL performs SVM-classification with all the tabulated features, SIFT histograms and twelve previously hand-picked concatenations of the tabulated features, selected on the basis of SOM classifier accuracy in the VOC2006 task. The runs HUTCIS_SVM_FULLIMG_IP+SC and HUTCIS_SVM_FULLIMG_IP are otherwise similar but use just subsets of the features: SIFT histograms and colour histogram, or just SIFT histograms, respectively.

The runs identified by prefix HUTCIS_SVM_BB are naive attempts to account for the different training and test image distributions. These runs are also based on SIFT histogram and colour histogram features. For the training images, the features are calculated from the bounding boxes specified in the VOC2006 annotations. For the test images, the features are calculated for whole images. The different runs with this prefix correspond to different ways of selecting the images as a basis for SIFT codebook formation.

The run HUTCIS_FULLIMG+BB is the rank based fusion of features extracted from full images and bounding boxes. The runs HUTCIS_PICSOM1 and HUTCIS_PICSOM2 are otherwise identical but use different settings of the SOM classifier parameters. HUTCIS_PICSOM2 smooths the feature spaces less, and the detection is based on more local information. Both the runs are based on the full set of features mentioned above.

Results: As expected, the run HUTCIS_SVM_FULLIMG_ALL with the full set of visual features extracted from the whole image turned out to be the best of our runs on average. However, for several individual query topics other runs produced better results. It remains unclear how much of the difference is explained by statistical fluctuations and how much by genuine differences between the various techniques on one hand, and between query topics on the other. However, by comparison with purely random AP values [10] it is reasonable to believe that some of the differences reflect real phenomena.

The mechanism for fusing the visual features was generic and straightforward. Still, using all of the features in a rather large set usually provided better performance than subsets of the features (HUTCIS_SVM_FULLIMG_ALL vs. HUTCIS_SVM_FULLIMG_IP+SC and HUTCIS_SVM_FULLIMG_IP), with some notable exceptions, especially query “motorbike”. This is in line with our general observation (and common knowledge) that without specific knowledge

of the target objects, an acceptable solution can often be found by blindly fusing a large number of features.

In general, it was found better to train with features extracted from whole images instead of just bounding boxes (e.g. HUTCIS_SVM_FULLIMG_IP+SC and HUTCIS_SVM_BB_BB_IP+SC), with possible exception in the query “person”. This is no surprise given the unsymmetry in our feature extraction and matching: the features extracted from bounding boxes of the training objects were compared with the features of all of the test images. The bounding box technique does not even seem to give much complementary information in addition to the full image information, as fusing these approaches (HUTCIS_SVM_FULLIMG+BB) usually results in worse performance than using the full images alone.

The results of the SOM classifier runs did not provide information that would be of general interest, besides confirming the previously known result of SOM classifiers being inferior to SVMs.

3.3 INAOE’s Annotation-Based Object Retrieval Approaches

Authors: Heidy Marisol Marin Castro, Hugo Jair Escalante Balderas, and Carlos Arturo Hernández Gracidas

Affiliation: TIA Research Group, Computer Science Department, National Institute of Astrophysics, Optics and Electronics, Tonantzintla, Mexico

Email: {hmarinc,hugojair,carloshg}@ccc.inaoep.mx

The TIA research group at INAOE, Mexico proposed two methods based on image labeling. Automatic image annotation methods were used for labeling regions within segmented images, and then we performed object retrieval based on the generated annotations. Two approaches were proposed: a semi-supervised classifier based on unlabeled data and a supervised one, where the latter method was enhanced by a recently proposed method based on semantic cohesion [20]. Both approaches followed the following steps:

1. Image segmentation
2. Feature extraction
3. Manual labeling of a small subset of the training set
4. Training a classifier
5. Using the classifier for labeling the test-images
6. Using labels assigned to region images for object retrieval

For both approaches the full collection of images was segmented with the normalized cuts algorithm [21]. A set of 30 features were extracted from each region; we considered color, shape and texture attributes. We used our own tools for image segmentation, feature extraction and manual labeling [22]. The considered annotations were the labels of the 10 objects defined for this task. The features for each region together with the manual annotations for each region were used as the training set with the two approaches proposed. Each classifier was trained with this dataset and then all of the test images were annotated

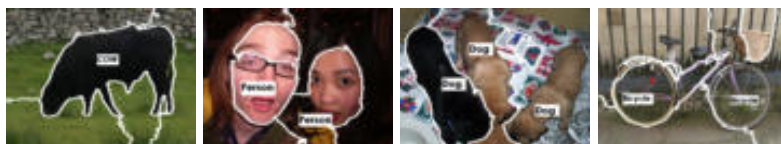


Fig. 4. Sample images from the generated training set

with such a classifier. Finally, the generated annotations were used for retrieving objects with queries. Queries were created using the labels of the objects defined for this task; and selected as relevant those images with the highest number of regions annotated with the object label. Sample segmented images with their corresponding manual annotations are shown in Figure 4. As we can see the segmentation algorithm works well for some images (isolated cows, close-up of people), however for other objects segmentation is poor (a bicycle, for example).

***KNN+MRFI*, A Supervised Approach:** For the supervised approach we used a simple *knn* classifier for automatically labeling regions. Euclidean distance was used as the similarity function. The label of the nearest neighbor (in the training set) for each test-region was assigned as annotation for this region. This was our baseline run (*INAOE-TIA-INAOE-RB-KNN*).

The next step consisted of improving the annotation performance of *knn* using an approach called *MRFI* [20] which we recently proposed for improving annotation systems. This approach consists of modeling each image (region-annotations pairs) with a Markov random field (*MRF*), introducing semantic knowledge, see Figure 5. The top- k more likely annotations for each region are considered. Each of these annotations has a confidence weight related to the relevance of the label to being the correct annotation for that region, according to *knn*. The *MRFI* approach uses the relevance weights with semantic information for choosing a unique (the correct) label for each region. Semantic information is considered in the *MRF* for keeping coherence among annotations assigned to regions within a common image; while the relevance weight is considered for taking into account the confidence of the annotation method ($k - nn$) on each of the labels, see Figure 5. The (pseudo) optimal configuration of region-annotations for each image is obtained by minimizing an energy function defined by potentials. For optimization we used standard simulated annealing.

The intuitive idea of the *MRFI* approach is to guarantee that the labels assigned to regions are coherent among themselves, taking into account semantic knowledge and the confidence of the annotation system. In previous work, semantic information was obtained from cooccurrences of labels on an external corpus. However for this work semantic association between a pair of labels is given by the normalized number of relevant documents returned by Google^R to queries generated using the pair of labels. This run is named *INAOE-TIA-INAOE-RB-KNN+MRFI*, see [20] for details.

SSAssemble: Semi-Supervised Weighted AdaBoost: The semi-supervised approach consists of using a recently proposed ensemble of classifiers, called WSA

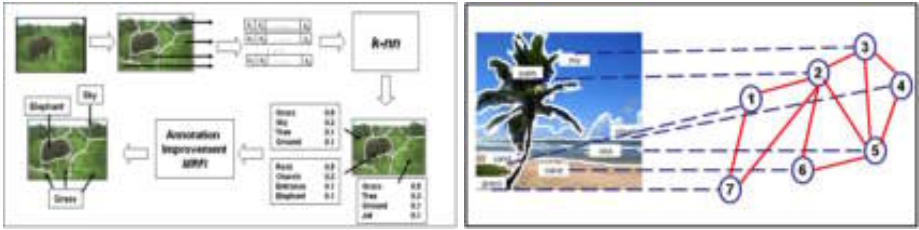


Fig. 5. Left: graphical description of the improvement process of *MRFI*. Right: interpretation of *MRFI* for a given configuration of labels and regions; (red) line-arcs consider semantic cohesion between labels, while (blue) dashed-arcs consider relevance weight of each label according to $k - nn$.

[22]. Our WSA ensemble uses naive Bayes as its base classifier. A set of these is combined in a cascade based on the AdaBoost technique [23]. Ensemble methods work by combining a set of base classifiers in some way, such as a voting scheme, producing a combined classifier which usually outperforms a single classifier. When training the ensemble of Bayesian classifiers, WSA considers the unlabeled images at each stage. These are annotated based on the classifier from the previous stage, and then used to train the next classifier. The unlabeled instances are weighted according to a confidence measure based on their predicted probability value; while the labeled instances are weighted according to the classifier error, as in standard AdaBoost. Our method is based on the supervised multi-class AdaBoost ensemble, which has shown to be an efficient scheme to reduce the error rate of different classifiers.

Formally the WSA algorithm receives a set of labeled data (L) and a set of unlabeled data (U). An initial classifier NB_1 is build using L . The labels in L are used to evaluate the error of NB_1 . As in AdaBoost the error is used to weight the examples, increasing the weight of the misclassified examples and keeping the same weight of the correctly classified examples. The classifier is used to predict a class for U with certain probability. In the case of U , the weights are multiplied by the predicted probability of the majority class. Unlabeled examples with high probability of their predicted class will have more influence in the construction of the next classifier than examples with lower probabilities. The next classifier NB_2 is build using the weights and predicted class of $L \cup U$. NB_2 makes new predictions on U and the error of NB_2 on all the examples is used to reweight the examples. This process continues, as in AdaBoost, for a predefined number of cycles or when a classifier has a weighted error greater than or equal to 0.5. As in AdaBoost, new instances are classified using a weighted sum of the predicted class of all the constructed base classifiers. WSA is described in Algorithm 2.

We faced several problems when performing the annotation image task. The first one was that the training set and the test set were different, so this caused a classification with high error ratio. The second one was due the segmentation algorithm. The automatic segmentation algorithm did not perform well for all images leading to incorrect segmentation of the objects in the images. The last one concerns the different criteria for manual labeling of the training set. Due to

all these facts we did not get good results. We hope to improve the annotation task by changing part of the labeling strategy.

3.4 MSRA: Object Retrieval

Authors: Mingjing Li, Xiaoguang Rui, and Lei Wu

Affiliation: Microsoft Research Asia

Email: mjli@microsoft.com

Two approaches were adopted by Microsoft Research Asia (MSRA) to perform the object retrieval task in ImageCLEF 2007. One is based on the visual topic model (VTM); the other is the visual language modelling (VLM) method [24]. VTM represents an image by a vector of probabilities that the image belongs to a set of visual topics, and categorizes images using SVM classifiers. VLM represents an image as a 2-D document consisting of visual words, trains a statistical language model for each image category, and classifies an image to the category that generates the image with the highest probability.

VTM: Visual Topic Model: Probabilistic Latent Semantic Analysis (pLSA) [25], which is a generative model from the text literature, is adopted to find out the visual topics from training images. Different from traditional pLSA, all training images of 10 categories are put together in the training process and about 100 visual topics are discovered finally.

The training process consists of five steps: local feature extraction, visual vocabulary construction, visual topic construction, histogram computation, and classifier training. At first, salient image regions are detected using scale invariant interest point detectors such as the Harris-Laplace and the Laplacian detectors. For each image, about 1,000 to 2,000 salient regions are extracted. Those regions are described by the SIFT descriptor which computes a gradient orientation histogram within the support region. Next, 300 local descriptors are randomly selected from each category and combined together to build a global vocabulary of 3,000 visual words. Based on the vocabulary, images are represented by the frequency of visual words. Then, pLSA is performed to discover the visual topics in the training images. pLSA is also applied to estimate how likely an image belongs to each visual topic. The histogram of the estimated probabilities is taken as the feature representation of that image for classification. For multi-class classification problem, we adopt the one-against-one scheme, and train an SVM classifier with RBF kernel for each possible pair of categories.

VLM: Visual Language Modeling: The approach consists of three steps: image representation, visual language model training and object retrieval. Each image is transformed into a matrix of visual words. First, an image is simply segmented into 8×8 patches, and the texture histogram feature is extracted from each patch. Then all patches in the training set are grouped into 256 clusters based on their features. Next, each patch cluster is represented using an 8-bit hash code, which is defined as the visual word. Finally, an image is represented by a matrix of visual words, which is called a *visual document*.

Algorithm 2. Semi-supervised Weighted AdaBoost (WSA) algorithm.

Require: L : labeled instances, U : unlabeled instances, P : training instances, T : Iterations

Ensure: Final Hypothesis and probabilities: $H_f = \operatorname{argmax}_{\sum_{t=1}^T \log \frac{1}{B_t}}, P(x_i)$

```

1:  $W(x_i)^0 = \frac{1}{\text{NumInst}(L)}, \forall x_i \in L$ 
2: for  $t$  from 1 to  $T$  do
3:    $W(x_i)^t = \frac{W(x_i)^{t-1}}{\sum_{i=1}^N W(x_i)^{t-1}} \forall x_i \in L$ 
4:    $h_t = C(L, W(x_i)^t)$ 
5:    $e_t = \sum_{i=1}^N W(x_i)^t$  if  $h_t(x_i) \neq y_i$ 
6:   if  $e_t \geq 0.5$  then
7:     exit
8:   end if
9:   if  $e_t = 0.0$  then
10:     $e_t = 0.01$ 
11:  end if
12:   $B_t = \frac{e_t}{(1-e_t)}$ 
13:   $W(x_i)^{(t+1)} = W(x_i)^t * B_t$  if  $h_t(x_i) = y_i \forall x_i \in L$ 
14:   $P(x_i) = C(L, U, W(x_i)^t)$ 
15:   $W(x_i) = P(x_i) * B_t \forall x_i \in U$ 
16: end for
```

Visual words in a visual document are not independent to each other, but correlated with other words. To simplify the model training, we assume that visual words are generated in the order from left to right, and top to bottom and each word is only conditionally dependent on its immediate top and left neighbors, and train a trigram language model for each image category. Given a test image, it is transformed into a matrix of visual words in the same way, and the probability that it is generated by each category is estimated respectively. Finally, the image categories are ranked in the descending order of these probabilities.

3.5 NTU: Solution for the Object Retrieval Task

Authors: Steven C. H. Hoi

Affiliation: School of Computer Engineering, Nanyang Technological University, Singapore

Email: chhoi@ntu.edu.sg

Introduction: Object retrieval is an interdisciplinary research problem between object recognition and content-based image retrieval (CBIR). It is commonly expected that object retrieval can be solved more effectively with the joint maximization of CBIR and object recognition techniques. We study a typical CBIR

solution with application to the object retrieval tasks [26,27]. We expect that the empirical study in this work will serve as a baseline for future research when using CBIR techniques for object recognition.

Overview of Our Solution: We study a typical CBIR solution for the object retrieval problem. In our approach, we focus on two key tasks. One is the feature representation, the other is the supervised learning scheme with support vector machines.

Feature Representation: In our approach, three kinds of global features are extracted to represent an image, including color, shape, and texture.

For color, we study the Grid Color Moment feature (GCM). We split each image into a 3×3 grid and extract color moments to represent each of the 9 regions of the grid. Three color moments are then computed: color mean, color variance and color skewness in each color channel (H, S, and V), respectively. Thus, an 81-dimensional color moment is adopted as the color feature for each image.

For shape, we employ the edge direction histogram. First, an input color image is converted into a grayscale image. Then a Canny edge detector is applied to obtain its edge image. Based on the edge images, the edge direction histogram can be computed. Each edge direction histogram is quantized into 36 bins of 10 degrees each. In addition, we use a bin to count the number of pixels without edges. Hence, a 37-dimensional edge direction histogram is used for shape.

For texture, we investigate the Gabor feature. Each image is first scaled to the size of 64×64 . Then, the Gabor wavelet transformation is applied to the scaled image at 5 scale levels and 8 orientations, which results in a total of 40 subimages for each input image. For each subimage, we calculate three statistical moments to represent the texture, including mean, variance, and skewness. Therefore, a 120-dimensional feature vector is used for texture.

In total, a 238-dimensional feature vector is used to represent each image. The set of visual features has been shown to be effective for content-based image retrieval in our previous experiments [26,27].

Supervised Learning for Object Retrieval: The object retrieval task defined in ImageCLEF 2007 is similar to a relevance feedback task in CBIR, in which a number of positive and negative labeled examples are given for learning. This can be treated as a supervised classification task. To solve it, we employ the support vector machines (SVM) technique for training the classifiers on the given examples [26]. In our experiment, a standard SVM package is used to train the SVM classifier with RBF kernels. The parameters C and γ are best tuned on the VOC 2006 training set, in which the training precision is 84.2% for the classification tasks. Finally, we apply the trained classifiers to do the object retrieval by ranking the distances of the objects from the classifier's decision boundary.

Concluding Remarks: We found that the current solution, though it was trained with good performance in an object recognition test-bed, did not achieve

promising results in the tough object retrieval tasks. In our future work, several directions can be explored to improve the performance, including local feature representation and better machine learning techniques.

3.6 PRIP: Color Interest Points and SIFT Features

Authors: Julian Stöttinger¹, Allan Hanbury¹, Nicu Sebe², Theo Gevers²

Affiliation: ¹ PRIP, Institute of Computer-Aided Automation, Vienna University of Technology, Vienna, Austria; ² Intelligent Systems Lab Amsterdam, University of Amsterdam, The Netherlands

Email: {julian,hanbury}@prip.tuwien.ac.at,
{nicu,gevers}@science.uva.nl

In the field of retrieval, detection, recognition and classification of objects, many state of the art methods use interest point detection at an early stage. This initial step typically aims to find meaningful regions in which descriptors are calculated. Finding salient locations in image data is crucial for these tasks. Most current methods use only the luminance information of the images. This approach focuses on the use of color information in interest point detection and its gain in performance. Based on the Harris corner detector, multi-channel visual information transformed into different color spaces is the basis to extract the most salient interest points. To determine the characteristic scale of an interest point, a global method of investigating the color information on a global scope is used. The two PRIP runs differ in the properties of these interest points only. The method consists of the following stages:

1. Extraction of multi-channel based interest points
2. Local descriptions of interest points
3. Estimating the signature of an image
4. Classification

Extraction of Multi-Channel Based Interest Points: An extension of the intensity-based Harris detector [28] is proposed in [29]. Because of common photometric variations in imaging conditions such as shading, shadows, specularities and object reflectance, the components of the *RGB* color system are correlated and therefore sensitive to illumination changes. However, in natural images, high contrast changes may appear. Therefore, a color Harris detector in *RGB* space does not dramatically change the position of the corners compared to a luminance based approach. Normalized *rgb* overcomes the correlation of *RGB* and favors color changes. The main drawback, however, is its instability in dark regions. We can overcome this by using quasi invariant color spaces.

The approach PRIP-PRIP_HSLScIvHarris uses the *HSI* color space [30], which is quasi-invariant to shadowing and specular effects. Therefore, changes in lighting conditions in images should not affect the positions of the interest points, resulting in more stable locations. Additionally, the *HSI* color space discriminates between luminance and color. Therefore, much information can be discarded, and the locations get more sparse and distinct.

The PRIP_cbOCS_ScIvHarris approach follows a different idea. As proposed in [31], colors have different occurrence probabilities and therefore different information content. Therefore, rare colors are regarded as more salient than common ones. We use a boosting function so that color vectors having equal information content have equal impact on the saliency function. This transformation can be found by analyzing the occurrence probabilities of colors in large image databases. With this change of focus towards rare colors, we aim to discard many repetitive locations and get more stable results on rare features.

The characteristic scale of an interest point is chosen by applying a principal component analysis (PCA) on the image and thus finding a description for the correlation of the multi-channel information [32]. The characteristic scale is decided when the Laplacian of Gaussian function of this projection and the Harris energy is a maximum at the same location in the image. The final extraction of these interest points and corresponding scales is done by preferring locations with high Harris energy and large scales. A maximum number of 300 locations per image has been extracted, as over-description diminishes the overall recognition ability.

Local Descriptions of Interest Points: The scale invariant feature transform (SIFT) [33] showed to give best results in a broad variety of applications [34]. We used the areas of the extracted interest points as a basis for the description phase. SIFT are basically sampled and normalized gradient histograms, which can lead to multiple descriptions per location. This occurs if there is more than one direction of the gradients regarded as predominant.

Estimating the Signature of an Image: In this bag of visual features approach [35], we cluster the descriptions of one image to a fixed number of 40 clusters using k-means. The centroids and the proportional sizes of the clusters build the signature of one image having a fixed dimensionality of 40 by 129.

Classification: The Earth Mover's Distance (EMD) [36] showed to be a suitable metric for comparing image signatures. It takes the proportional sizes of the clusters into account, which gains much discriminative power. The classification itself is done in the most straightforward way possible: for every object category, the smallest distances to another signature indicate the classification.

3.7 RWTHi6: Patch-Histograms and Log-Linear Models

Authors: Thomas Deselaers, Hermann Ney

Affiliation: Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany

Email: surname@cs.rwth-aachen.de

The approach used by the Human Language Technology and Pattern Recognition group of the RWTH Aachen University, Aachen, Germany, to participate in the PASCAL Visual Object Classes Challenge consists of four steps:

1. patch extraction
2. clustering

3. creation of histograms
4. training of a log-linear model

where the first three steps are feature extraction steps and the last is the actual classification step. This approach was first published in [37,38].

The method follows the promising approach of considering objects to be constellations of parts which offers the immediate advantages that occlusions can be handled very well, that the geometrical relationship between parts can be modelled (or neglected), and that one can focus on the discriminative parts of an object. That is, one can focus on the image parts that distinguish a certain object from other objects.

The steps of the method are briefly outlined in the following paragraphs. To model the difference in the training and test data, the first three steps have been done for the training and test data individually, and then the corresponding histograms have been extracted for the respective other, so that the vocabulary was learnt once for the training data and once for the test data, and the histograms are created for each using both vocabularies. Results however show that this seems not to be a working approach to tackle divergence in training and testing data.

Patch Extraction: Given an image, we extract square image patches at up to 500 image points. Additionally, 300 points from a uniform grid of 15×20 cells that is projected onto the image are used. At each of these points a set of square image patches of varying sizes (in this case 7×7 , 11×11 , 21×21 , and 31×31 pixels) are extracted and scaled to a common size (in this case 15×15 pixels).

In contrast to the interest points from the detector, the grid-points can also fall onto very homogeneous areas of the image. This property is on the one hand important for capturing homogeneity in objects which is not found by the interest point detector and on the other hand it captures parts of the background which usually is a good indicator for an object, as in natural images objects are often found in a “natural” environment.

After the patches are extracted and scaled to a common size, a PCA dimensionality reduction is applied to reduce the large dimensionality of the data, keeping 39 coefficients corresponding to the 40 components of largest variance but discarding the first coefficient corresponding to the largest variance. The first coefficient is discarded to achieve a partial brightness invariance. This approach is suitable because the first PCA coefficient usually accounts for global brightness.

Clustering: The data are then clustered using a k -means style iterative splitting clustering algorithm to obtain a partition of all extracted patches. To do so, first one Gaussian density is estimated which is then iteratively split to obtain more densities. These densities are then re-estimated using k -means until convergence is reached and then the next split is done. It has been shown experimentally that results consistently improve up to 4096 clusters but for more than 4096 clusters the improvement is so small that it is not worth the higher computational demands.

Creation of Histograms: Once we have the cluster model, we discard all information for each patch except its closest corresponding cluster center identifier. For the test data, this identifier is determined by evaluating the Euclidean distance to all cluster centers for each patch. Thus, the clustering assigns a cluster $c(x) \in \{1, \dots, C\}$ to each image patch x and allows us to create histograms of cluster frequencies by counting how many of the extracted patches belong to each of the clusters. The histogram representation $h(X)$ with C bins is then determined by counting and normalization such that $h_c(X) = \frac{1}{L_X} \sum_{l=1}^{L_X} \delta(c, c(x_l))$, where δ denotes the Kronecker delta function, $c(x_l)$ is the closest cluster center to x_l , and x_l is the l -th image patch extracted from image X , from which a total of L_X patches are extracted.

Training & Classification: Having obtained this representation by histograms of image patches, we define a decision rule for the classification of images. The approach based on maximum likelihood of the class-conditional distributions does not take into account the information of competing classes during training. We can use this information by maximizing the class posterior probability $\prod_{k=1}^K \prod_{n=1}^{N_k} p(k|X_{kn})$ instead. Assuming a Gaussian density with pooled covariances for the class-conditional distribution, this maximization is equivalent to maximizing the parameters of a log-linear or maximum entropy model

$$p(k|h) = \frac{1}{Z(h)} \exp \left(\alpha_k + \sum_{c=1}^C \lambda_{kc} h_c \right),$$

where $Z(h) = \sum_{k=1}^K \exp \left(\alpha_k + \sum_{c=1}^C \lambda_{kc} h_c \right)$ is the renormalization factor. We use a modified version of generalized iterative scaling. Bayes' decision rule is used for classification.

4 Results

The results of this task published in [9] were shown to have several problems due to unclear relevance judgement guidelines and invalid submission files (e.g. wrong query order) [10].

Therefore a thorough analysis of all submitted runs was performed for this work and the results presented here differ in part significantly from those presented in [9]. In particular,

- all runs were carefully checked to fully comply with the latest version of `trec_eval` and to deliver a maximum of 1,000 results per class;
- based on the full annotation of the database by Ville Viitaniemi [10], the pooling was re-done and new relevance judgements were created as they would have been if judging guidelines would have been more clear and all runs would have had proper formatting.

The results presented here are all fully comparable except for the two runs from the Budapest group. They assigned one class-label per image instead of possibly

Table 4. Results from the ImageCLEF 2007 object retrieval task using the relevance judgements obtained from simulated pooling. All values have been multiplied by 100 to make the table more readable. The numbers in the top row refer to the class id’s (see Table 1). The MAP over all classes is in the last column. The highest AP per class is shown in bold.

run id	query										MAP
	1	2	3	4	5	6	7	8	9	10	
HUTCIS_SVM_FULLIMG_ALL	18.7	4.0	22.7	1.7	0.0	2.4	0.8	13.1	0.0	18.5	9.1
HUTCIS_SVM_FULLIMG_IP+SC	9.1	2.9	21.7	4.3	0.0	4.1	1.4	11.6	0.0	18.8	8.2
HUTCIS_SVM_FULLIMG_IP	8.2	3.1	20.2	8.6	0.0	4.6	0.7	11.5	0.0	16.2	8.1
HUTCIS_SVM_FULLIMG+BB	12.1	3.7	9.5	2.7	0.0	2.4	2.1	8.7	0.0	20.7	6.9
HUTCIS_SVM_BB_ALL	6.0	3.3	1.7	1.4	0.0	2.2	0.8	6.0	0.0	22.7	4.9
HUTCIS_SVM_BB_BB_IP+SC	5.2	3.3	2.2	1.6	0.0	1.5	0.4	4.0	0.0	22.5	4.5
HUTCIS_SVM_BB_FULL_IP+SC	8.3	2.3	1.0	0.9	0.0	2.6	0.4	4.0	0.0	20.7	4.5
HUTCIS_SVM_BB_BAL_IP+SC	4.8	2.7	1.7	1.0	0.0	1.5	0.9	2.5	0.0	22.4	4.2
HUTCIS_SVM_BB_BB_IP	3.9	1.6	0.9	7.0	0.0	1.0	0.3	2.8	0.0	16.9	3.8
HUTCIS_PICSOM1	2.9	2.4	12.3	2.0	0.0	0.9	0.7	1.2	0.0	10.4	3.6
HUTCIS_SVM_BB_BAL_IP	3.8	2.2	0.7	1.5	0.0	0.9	0.8	2.5	0.0	17.7	3.3
HUTCIS_PICSOM2	1.6	2.3	12.1	1.6	0.0	0.6	1.0	0.8	0.0	9.1	3.2
MSRA-MSRA_RuiSp	2.7	1.4	7.5	2.4	2.3	0.1	0.9	0.4	0.0	10.6	3.1
HUTCIS_SVM_BB_FULL_IP	0.4	2.7	0.5	1.3	0.0	0.9	0.3	2.9	0.0	13.9	2.5
NTU_SCE_HOI-NTU_SCE_HOI_1	4.2	2.0	4.5	0.0	0.0	0.0	0.3	0.1	0.0	0.2	1.3
RWTHi6-HISTO-PASCAL	0.4	0.4	1.3	0.6	0.0	0.1	0.0	0.2	0.0	7.1	1.1
budapest-acad-budapest-acad314	0.4	0.1	1.2	0.0	0.0	0.6	0.5	0.6	0.0	5.7	1.0
budapest-acad-budapest-acad315	2.0	0.0	0.6	0.5	0.0	0.0	0.0	0.0	0.0	5.3	1.0
PRIP-PRIP_HSI_ScIvHarris	0.3	0.0	0.4	0.1	4.6	0.5	0.0	0.0	0.0	1.5	0.8
MSRA-MSRA-VLM-8-8-640-ful	0.7	0.6	0.8	0.3	0.0	0.0	0.2	0.8	0.0	2.6	0.7
MSRA-MSRA-VLM-8-8-800-HT	1.1	0.4	0.7	0.2	0.0	0.0	0.0	0.5	0.0	2.2	0.6
INAOE-TIA-INAOE-RB-KNN+MRFI	0.7	0.0	0.5	0.0	0.0	0.6	0.0	0.0	0.0	2.5	0.5
INAOE-TIA-INAOE-RB-KNN+MRFI_ok	0.7	0.0	0.5	0.0	0.0	0.6	0.0	0.0	0.0	2.5	0.5
INAOE-TIA-INAOE-RB-KNN	0.0	0.0	0.3	0.2	0.0	0.0	0.0	0.0	0.0	3.3	0.4
PRIP-PRIP_cbOCS_ScIvHarr2	0.1	0.0	0.1	1.5	0.2	0.0	0.0	0.0	0.0	0.6	0.3
INAOE-TIA-INAOE_SSAssemble	0.4	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.6	0.2

several ones (e.g. there may be a bicycle and a person in an image). Furthermore they used different, more strongly labelled training data.

Table 4 gives results for all runs using the relevance judgements obtained from simulated pooling and Table 5 gives the same results but uses the relevance information for the whole database. The tables are ordered by MAP (last column). The ordering, however should not be interpreted as a general ranking of the methods since the methods perform very differently among the different topics.

5 Discussion

In this section, the results for the full database annotation are discussed in more detail. However most of the observations can also be found in the results obtained using the simulated pooling.

Considering the class-wise results, it can be observed that the best overall results were obtained for the *car* query (column 3), for which the best run has an AP of about 11%. This can clearly be useful in a practical application. The best run for the *bicycle* class (column 1) is also able to find enough relevant images to be useful in a practical application.

Table 5. Results from the ImageCLEF 2007 object retrieval task with complete relevance information for the whole database. All values have been multiplied by 100 to make the table more readable. The numbers in the top row refer to the class id's (see Table 1). The MAP over all classes is in the last column. The highest AP per class is shown in bold.

	run id	query										MAP
		1	2	3	4	5	6	7	8	9	10	
HUTCIS_SVM_FULLIMG_ALL	4.1	1.2	10.6	0.4	0.0	0.6	0.1	3.8	0.0	8.3	2.9	
HUTCIS_SVM_FULLIMG_IP+SC	2.6	1.0	11.1	1.0	0.0	1.0	0.1	3.2	0.0	8.2	2.8	
HUTCIS_SVM_FULLIMG_IP	2.4	1.1	10.3	1.8	0.0	1.1	0.1	3.0	0.0	8.1	2.8	
HUTCIS_SVM_FULLIMG+BB	3.0	1.1	4.2	0.6	0.0	0.7	0.1	2.5	0.0	8.6	2.1	
HUTCIS_SVM_BB_ALL	1.6	0.9	0.5	0.3	0.0	0.6	0.1	1.5	0.0	8.3	1.4	
HUTCIS_SVM_BB_BB_IP+SC	1.4	1.0	0.7	0.3	0.0	0.5	0.1	1.1	0.0	8.4	1.4	
HUTCIS_SVM_BB_FULL_IP+SC	2.0	0.8	0.4	0.2	0.0	0.8	0.1	1.1	0.0	8.2	1.3	
HUTCIS_PICSOM1	0.9	0.7	4.5	0.6	0.0	0.3	0.1	0.7	0.0	5.6	1.3	
MSRA-MSRA_RuiSp	0.9	0.5	3.6	0.6	0.7	0.1	0.1	0.4	0.0	6.0	1.3	
HUTCIS_SVM_BB_BAL_IP+SC	1.3	0.8	0.5	0.2	0.0	0.5	0.1	0.8	0.0	8.4	1.3	
HUTCIS_PICSOM2	0.8	0.6	4.2	0.5	0.0	0.3	0.1	0.4	0.0	5.4	1.2	
HUTCIS_SVM_BB_BB_IP	1.1	0.7	0.4	1.4	0.0	0.3	0.0	1.0	0.0	7.2	1.2	
HUTCIS_SVM_BB_BAL_IP	1.1	0.8	0.3	0.3	0.0	0.4	0.1	0.9	0.0	6.9	1.1	
HUTCIS_SVM_BB_FULL_IP	0.3	0.9	0.3	0.3	0.0	0.3	0.0	1.1	0.0	6.6	1.0	
RWTHi6-HISTO-PASCAL	0.4	0.2	1.4	0.2	0.0	0.1	0.0	0.2	0.0	5.5	0.8	
budapest-acad-budapest-acad314	0.1	0.1	0.8	0.0	0.0	0.2	0.0	0.2	0.0	4.1	0.5	
NTU_SCE_HOI-NTU_SCE_HOI	1.2	0.7	2.4	0.0	0.0	0.0	0.1	0.1	0.0	0.8	0.5	
budapest-acad-budapest-acad315	0.4	0.0	0.4	0.1	0.0	0.0	0.1	0.1	0.0	3.9	0.5	
MSRA-MSRA-VLM_8_8_640_ful	0.4	0.3	0.7	0.1	0.1	0.0	0.0	0.3	0.0	2.5	0.4	
MSRA-MSRA-VLM-8-8-800-HT	0.3	0.2	0.5	0.0	0.0	0.1	0.0	0.2	0.1	2.5	0.4	
INAOE-TIA-INAOE_SSAssemble	0.1	0.0	0.1	0.0	0.0	0.2	0.0	0.2	0.0	3.2	0.4	
INAOE-TIA-INAOE-RB-KNN+MRFI	0.5	0.1	0.6	0.0	0.0	0.2	0.0	0.0	0.0	2.2	0.4	
INAOE-TIA-INAOE-RB-KNN+MRFI _{ok}	0.5	0.1	0.6	0.0	0.0	0.2	0.0	0.0	0.0	2.2	0.4	
PRIP-PRIP_HSL_SclvHarris	0.1	0.0	0.3	0.1	1.4	0.1	0.0	0.0	0.0	1.5	0.4	
INAOE-TIA-INAOE-RB-KNN	0.3	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	2.2	0.3	
PRIP-PRIP _{cb} OCS_SclvHarr2	0.1	0.0	0.1	0.5	0.1	0.0	0.1	0.1	0.0	0.8	0.2	

It is also clear that the three classes for which the best APs are significantly high (car, bicycle and person) are also those classes with the highest number of relevant images (Table 1). This could be because having a high number of relevant images in the dataset means that they have a higher probability of being detected by chance. However, the bad performance on these classes by some of the methods also provides evidence against this conjecture.

The results for the classes having fewer relevant images in the dataset are less easy to interpret and generalise. For the *bus* class (column 2), although the better runs are able to find a few images showing buses, these images are not ranked very highly. By joining all runs, only 140 of the 218 images in the database that show buses are found. The best run for the *cat* class (column 5) obtains an average precision of 1.4% which can already be considered a promising result given the extremely low number of relevant images in the database. The best run for the *cow* query (column 6) finds 12 out of 49 relevant images. However, only one of the images is among the top 10 retrieved. For queries 7 (*dog*) and 9 (*sheep*), the best runs have an AP of 0.1%, which is not high enough for use in a practical application. The best result for the *horse* query (column 8) finds 49 of the 175 relevant images in the database.

The *person* query is certainly to be treated differently than all other queries since the number of relevant images for this query is higher than the allowed number of results returned. The best run returns 984 images showing persons which is 98.4% of the optimal result. Several other runs find more than 800 relevant images. However, all runs jointly only found 6029 relevant images which is an indicator that most runs found similar, and probably “easier” images. While these algorithms produce promising results, they are not suitable for finding all images showing a person.

One issue that should be taken into account when interpreting the results is that about 50% of the evaluated runs are from HUTCIS and thus this group had a significant impact on the pools for relevance assessment. In the initial evaluation, this effect was further boosted by the fact that the initial runs from HUTCIS (which were used for the pooling) had the queries 4–10 in wrong order. This problem, was fixed in the evaluation here by simulating proper pooling using the annotation of the complete database. However, it can still be observed that the high number of HUTCIS runs makes them appear slightly better in the pooled results than in the results using the full database annotation. Additionally, we evaluated all runs with a different pooling strategy: the pooling was simulated with only one run per group, which removes the bias introduced by strongly differing numbers of submissions. Here we observed a ranking that is more similar to the ranking obtained when the annotation of the full database is used.

By comparing the results with and without pooling, it can be observed that pooling changes the results, however using the additional relevance information obtained during judging the pools, a more stable result can be obtained. The effect of pooling is particular strong for runs with only very few relevant images and for runs with very many relevant images.

The results clearly show that the task is a very difficult one and that it is very important to clearly define judging criteria and relevance assessment methods before running the evaluation. In particular it seems to be important to ensure an appropriate (not too few, not too many) number of relevant images per topic.

6 Conclusion

We presented the object retrieval task of ImageCLEF 2007, the methods of the participating groups, and the results. The main challenges inherent in the task were the difference in the nature of the images used for training and testing, as well as the large variation in the number of relevant images for each query, ranging from 6 to 11,248 of 20,000 images. The results show that none of the methods really solves the assigned task. Although large advances in object detection and recognition were achieved over the last years, still many improvements are necessary to solve difficult tasks with a high variability and only a restricted amount of training data. It can however be observed that some of the methods are able to obtain reasonable results for a limited set of classes. An interesting

observation is that few participating groups attempted to compensate for the differences in training and testing data, while the few attempts made were in general not successful.

Furthermore, the analysis of the results showed that the use of pooling techniques for relevance assessment can be problematic if the pools are biased due to erroneous runs or due to many strongly correlated submissions as it was the case in this evaluation.

Acknowledgements

We would like to thank the CLEF campaign for supporting the ImageCLEF initiative. This work was partially funded by the European Commission MUSCLE NoE (FP6-507752) and the DFG (German research foundation) under contract Ne-572/6.

We would like to thank the PASCAL NoE for allowing us to use the training data of the PASCAL 2006 Visual object classes challenge as training data for this task, in particular, we would like to thank Mark Everingham for his cooperation.

We would like to thank Paul Clough from Sheffield University for support in creating the pools and Jan Hosang, Jens Forster, Pascal Steingrube, Christian Plahl, Tobias Gass, Daniel Stein, Morteza Zahedi, Richard Zens, Yuqi Zhang, Markus Nußbaum, Gregor Leusch, Michael Arens, Lech Szumilas, Jan Bungeroth, David Rybach, Peter Fritz, Arne Mauser, Saša Hasan, and Stefan Hahn for helping to create relevance assessments for the images.

The work of the Budapest group was supported by a Yahoo! Faculty Research Grant and by grants *MOLINGV* NKFP-2/0024/2005, NKFP-2004 project Language Miner.

References

1. Everingham, M., et al.: The 2005 PASCAL Visual Object Classes Challenge. In: Quiñero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 117–176. Springer, Heidelberg (2006)
2. Everingham, M., Zisserman, A., Williams, C., Gool, L.V.: The Pascal Visual Object Classes Challenge 2006 (VOC2006) results. Technical report (2006), <http://www.pascal-network.org/>
3. Clough, P.D., Müller, H., Sanderson, M.: Overview of the CLEF cross-language image retrieval track (ImageCLEF) 2004. 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: Result of the fifth CLEF evaluation campaign, Bath, England. LNCS. Springer, Heidelberg (2005)
4. Moellic, P.A., Fluhr, C.: ImageEVAL 2006 official campaign. Technical report, ImagEVAL (2006)
5. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)

6. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
7. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The IAPR benchmark: A new evaluation resource for visual information systems. In: LREC 2006 OntoImage 2006: Language Resources for Content-Based Image Retrieval, Genoa, Italy (in press, 2006)
8. Braschler, M., Peters, C.: CLEF methodology and metrics. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 394–404. Springer, Heidelberg (2002)
9. Deselaers, T., Hanbury, A., Viitaniemi, V., Benczúr, A., Brendel, M., Daróczy, B., Escalante Balderas, H.J., Gevers, T., Hernández Gracidas, C.A., Hoi, S.C.H., Laaksonen, J., Li, M., Marin Castro, H.M., Ney, H., Rui, X., Sebe, N., Stöttinger, J., Wu, L.: Overview of the ImageCLEF 2007 object retrieval task. In: Working notes of the CLEF 2007 Workshop, Budapest, Hungary (2007)
10. Viitaniemi, V., Laaksonen, J.: Thoughts on evaluation of image retrieval inspired by ImageCLEF 2007 object retrieval task. In: MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation, Budapest, Hungary (2007)
11. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research* 5, 913–939 (2004)
12. Prasad, B.G., Biswas, K.K., Gupta, S.K.: Region-based image retrieval using integrated color, shape, and location index. *Computer Vision and Image Understanding* 94(1-3), 193–233 (2004)
13. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. *PAMI* 24(8), 1026–1038 (2002)
14. Lv, Q., Charikar, M., Li, K.: Image similarity search with compact data structures. In: CIKM 2004: Proceedings of the thirteenth ACM international conference on Information and knowledge management, pp. 208–217. ACM Press, New York (2004)
15. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. *International Journal of Computer Vision* 59 (2004)
16. Shi, J., Malik, J.: Normalized cuts and image segmentation. *PAMI* 22, 888–905 (2000)
17. Viitaniemi, V., Laaksonen, J.: Improving the accuracy of global feature fusion based image categorisation. In: Falcidieno, B., Spagnuolo, M., Avrithis, Y., Kompatsiaris, I., Buitelaar, P. (eds.) SAMT 2007. LNCS, vol. 4816, pp. 1–14. Springer, Heidelberg (2007)
18. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
19. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing* 13(4), 841–853 (2002)
20. Escalante, H.J., y Gómez, M.M., Sucar, L.E.: Word co-occurrence and MRFs for improving automatic image annotation. In: Proceedings of the 18th British Machine Vision Conference (BMVC 2007), Warwick, UK (September, 2007)
21. Shi, J., Malik, J.: Normalized cuts and image segmentation. *PAMI* 22(8), 888–905 (2000)

22. Marin-Castro, H.M., Sucar, L.E., Morales, E.F.: Automatic image annotation using a semi-supervised ensemble of classifiers. In: 12th Iberoamerican Congress on Pattern Recognition CIARP 2007, Viña del Mar, Valparaiso, Chile. LNCS. Springer, Heidelberg (to appear, 2007)
23. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning, pp. 148–156 (1996)
24. Wu, L., Li, M.J., Li, Z.W., Ma, W.Y., Yu, N.H.: Visual language modeling for image classification. In: 9th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR 2007), Augsburg, Germany (2007)
25. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: International Conference on Computer Vision, Beijing, China (2005)
26. Hoi, S.C.H., Lyu, M.R.: A novel log-based relevance feedback technique in content-based image retrieval. In: 12th ACM International Conference on Multimedia (MM 2004), New York, NY, USA, pp. 24–31 (2004)
27. Hoi, S.C., Lyu, M.R., Jin, R.: A unified log-based relevance feedback scheme for image retrieval. *IEEE Transactions on Knowledge and Data Engineering* 18(4), 509–524 (2006)
28. Harris, C., Stephens, M.: A combined corner and edge detection. In: 4th Alvey Vision Conference, pp. 147–151 (1988)
29. Montesinos, P., Gouet, V., Deriche, R.: Differential invariants for color images. In: ICPR, p. 838 (1998)
30. van de Weijer, J., Gevers, T.: Edge and corner detection by photometric quasi-invariants. *PAMI* 27(4), 625–630 (2005)
31. van de Weijer, J., Gevers, T., Bagdanov, A.: Boosting color saliency in image feature detection. *PAMI* 28(1), 150–156 (2006)
32. Stöttinger, J., Hanbury, A., Sebe, N., Gevers, T.: Do colour interest points improve image retrieval? In: ICIP (to appear, 2007)
33. Lowe, D.: Distinctive image features from scale-invariant keypoints. *IJCV* 60(2), 91–110 (2004)
34. Mikolaczyk, K., Schmid, C.: A performance evaluation of local descriptors. *PAMI* 27(10), 1615–1630 (2005)
35. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. *CWPR* 73(2), 213–238 (2006)
36. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. *IJCV* 40(2), 99–121 (2000)
37. 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)
38. Deselaers, T., Keysers, D., Ney, H.: Improving a discriminative approach to object recognition using image patches. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 326–333. Springer, Heidelberg (2005)

Overview of the ImageCLEFmed 2007 Medical Retrieval and Medical Annotation Tasks

Henning Müller^{1,2}, Thomas Deselaers³, Thomas M. Deserno⁴,
Jayashree Kalpathy-Cramer⁵, Eugene Kim⁵, and William Hersh⁵

¹ Medical Informatics, University and Hospitals of Geneva, Switzerland

² Business Information Systems, University of Applied Sciences Sierre, Switzerland

³ Computer Science Dep., RWTH Aachen University, Germany

⁴ Dept. of Medical Informatics, RWTH Aachen University, Germany

⁵ Oregon Health and Science University (OHSU), Portland, OR, USA

henning.mueller@sim.hcuge.ch

Abstract. This paper describes the medical image retrieval and medical image annotation tasks of ImageCLEF 2007. Separate sections describe each of the two tasks, with the participation and an evaluation of major findings from the results of each given. A total of 13 groups participated in the medical retrieval task and 10 in the medical annotation task.

The medical retrieval task added two new data sets for a total of over 66'000 images. Topics were derived from a log file of the Pubmed biomedical literature search system, creating realistic information needs with a clear user model.

The medical annotation task was in 2007 organized in a new format as a hierarchical classification had to be performed and classification could be stopped at any hierarchy level. This required algorithms to change significantly and to integrate a confidence level into their decisions to be able to judge where to stop classification to avoid making mistakes in the hierarchy. Scoring took into account errors and unclassified parts.

1 Introduction

ImageCLEF¹ [1,2] started within CLEF² (Cross Language Evaluation Forum [3]) in 2003 with the goal to benchmark image retrieval in multilingual document collections. A medical image retrieval task was added in 2004 to explore domain-specific multilingual information retrieval and also multi-modal retrieval by combining visual and textual features for retrieval. Since 2005, a medical retrieval and a medical image annotation task have both been part of ImageCLEF [4].

The important participation in CLEF and particularly ImageCLEF has shown the need for benchmarks, and their usefulness to the research community. In 2007, a total of 50 groups registered for ImageCLEF to get access to the data sets and tasks. Among these, 13 participated in the medical retrieval task and 10 in the medical automatic annotation task.

¹ <http://www.imageclef.org/>

² <http://www.clef-campaign.org/>

Other important benchmarks in the field of visual information retrieval include TRECVID³ on the evaluation of video retrieval systems [5], ImageEval⁴, mainly on visual retrieval of images and image classification, and INEX⁵ (Initiative for the Evaluation of XML retrieval) concentrating on retrieval of multimedia based on structured data. Close contact with these initiatives exists to develop complementary evaluation strategies.

This article focuses on the two medical tasks of ImageCLEF 2007, whereas two other papers [6,7] describe the new object classification task and the photographic retrieval task. More detailed information can also be found on the task web pages. An even more detailed analysis of the 2005 medical image retrieval task and its outcomes is also available in [8].

2 The Medical Image Retrieval Task

The medical image retrieval task has been run for four consecutive years. In 2007, two new databases were added for a total of more than 66'000 images in the collection. For the generation of realistic topics or information needs, log files of the medical literature search system Pubmed were used.

2.1 General Overview

Again and as in previous years, the medical retrieval task showed to be popular among research groups registering for CLEF in 2007. In total 31 groups from all continents and 25 countries registered. A total of 13 groups finally submitted 149 runs that were used for the pooling required for the relevance judgments.

2.2 Databases

In 2007, the same four datasets were used as in 2005 and 2006 and two new datasets were added. The *Casimage* dataset was made available to participants [9], containing almost 9'000 images of 2'000 cases [10]. Images present in Casimage included mostly radiology modalities, but also photographs, PowerPoint slides, and illustrations. Cases were mainly in French, with around 20% being in English and 5% without any annotation. We also used the *PEIR*⁶ (Pathology Education Instructional Resource) database with annotation based on the *HEAL*⁷ project (Health Education Assets Library, mainly Pathology images [11]). This dataset contained over 33'000 images with English annotations, with the annotation being on a per image and not a per case basis as in Casimage. The nuclear medicine database of MIR, the Mallinkrodt Institute of Radiology⁸ [12], was also

³ <http://www-nlpir.nist.gov/projects/t01v/>

⁴ <http://www.imageval.org/>

⁵ <http://inex.is.informatik.uni-duisburg.de/2006/>

⁶ <http://peir.path.uab.edu/>

⁷ <http://www.healcentral.com/>

⁸ <http://gamma.wustl.edu/home.html>

made available. This dataset contained over 2'000 images mainly from nuclear medicine with annotations provided per case and in English. The PathoPic⁹ collection (Pathology images [13]) was included in our dataset containing about 7'800 images, with extensive annotation on a per image basis in German. Part of the German annotation was translated into English.

In 2007, we added two new datasets. The first was the *myPACS*¹⁰ dataset of 15'140 images and 3'577 cases, all in English and containing mainly radiology images. The second was the Clinical Outcomes Research Initiative (*CORI*¹¹) Endoscopic image database containing 1'496 images with an English annotation per image and not per case. The latter database extended the spectrum of the total dataset since there were previously only a few endoscopic images in the dataset. An overview of all datasets is shown in Table 1.

Table 1. The databases used in ImageCLEFmed 2007

Collection Name	Cases	Images	Annotations	Annotations by Language
Casimage	2076	8725	2076	French – 1899, English – 177
MIR	407	1177	407	English – 407
PEIR	32319	32319	32319	English – 32319
PathoPIC	7805	7805	15610	German – 7805, English – 7805
myPACS	3577	15140	3577	English – 3577
Endoscopic	1496	1496	1496	English – 1496
Total	47680	66662	55485	French – 1899, English – 45781, German – 7805

2.3 Registration and Participation

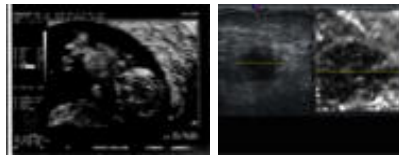
In 2007, 31 groups from all 6 continents and 25 countries registered for the ImageCLEFmed retrieval task, underlining the strong interest in this evaluation campaign. As in previous years, about half of the registered groups submitted results, with those not submitting results blaming a lack of time. The feedback from the non-submitting groups remains positive as they report that the data is a very useful resource. The following groups submitted results:

- CINDI group, Concordia University, Montreal, Canada;
- Dokuz Eylul University, Izmir, Turkey;
- IPAL/CNRS joint lab, Singapore, Singapore;
- IRIT–Toulouse, Toulouse, France;
- MedGIFT group, University and Hospitals of Geneva, Switzerland;
- Microsoft Research Asia, Beijing, China;
- MIRACLE, Spanish University Consortium, Madrid, Spain;
- MRIM–LIG, Grenoble, France;
- OHSU, Oregon Health & Science University, Portland, OR, USA;

⁹ <http://alf3.urz.unibas.ch/pathopic/intro.htm>

¹⁰ <http://www.mypacs.net/>

¹¹ <http://www.corii.org>



Ultrasound with rectangular sensor.

Ultraschallbild mit rechteckigem Sensor.

Ultrason avec capteur rectangulaire.

Fig. 1. Example for a visual topic

- RWTH Aachen Pattern Recognition group. Aachen, Germany;
- SINAI group, University of Jaen Intelligent Systems, Jaen, Spain;
- State University New York (SUNY) at Buffalo, NY, USA;
- UNAL group, Universidad Nacional Colombia, Bogotá, Colombia;

In total, 149 runs were submitted, with individual groups submitting anywhere from 1 to 36 runs. Several submitted runs had incorrect formats. These runs were corrected by the organizers whenever possible but a few runs were finally omitted from the pooling process and the final evaluation because `trec_eval` could not parse the results even after our modifications. Groups were able to re-score these runs as the `qrels` files were made available.

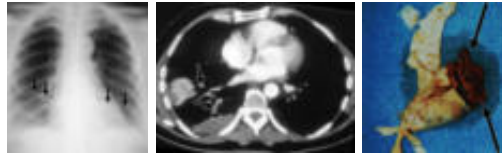
2.4 Query Topics

Query topics for 2007 were generated based on a log file of Pubmed¹². The log file of 24 hours contained a total of 77'895 queries. In general, the search terms were fairly vague and did not contain many image-related topics, so we filtered queries that had words such as image, video, and terms relating to modalities such as x-ray, CT, MRI, endoscopy, etc. We also aimed for the resulting terms to cover at least two or more of the following axes: modality, anatomic region, pathology, and visual observation (e.g., enlarged heart).

A total of 50 candidate topics were taken from these and sometimes an additional axis such as modality was added. From these topics we checked whether at least a few relevant images were in the database and from this, 30 topics were selected.

All topics were categorized with respect to the retrieval approach expected to perform best: visual topics, textual (semantic) topics and mixed topics. This was performed by an experienced image retrieval system developer. For each of the three retrieval approaches, 10 topics were selected for a total of 30 query topics that were distributed among the participants. Each topic consisted of the query itself in three languages (English, German, French) and 2–3 example images for the visual retrieval. Topic images were obtained from the Internet and were not part of the database. This made visual retrieval hard as most images were taken from different collections than those in the database and had changes in the gray level or color values.

¹² <http://www.pubmed.gov/>



Pulmonary embolism all modalities.
 Lungenembolie alle Modalitäten.
 Embolie pulmonaire, toutes les formes.

Fig. 2. Example for a semantic topic

Figure 1 shows a visual topic, and Figure 2 a topic with very different images in the results sets that should be well-suited for textual retrieval, only.

2.5 Relevance Judgments

Relevance judgments were performed by physicians who were students in the OHSU biomedical informatics graduate program. All were paid an hourly rate for their work. The pools for relevance judging were created by selecting the top ranking images from all submitted runs. The actual number selected from each run has varied by year. In 2007, it was 35 images per run, with the goal of having pools of about 800-1200 images in size for judging. The average pool size in 2007 was 890 images. Judges were instructed to rate images in the pools as definitely relevant (DR), partially relevant (PR), or not relevant (NR). Judges were instructed to use the partially relevant designation only in case they could not determine whether the image in question was relevant.

One of the problems was that all judges were English speakers but that the collection had a fairly large number of French and German documents. If the judgment required reading the text, judges had more difficulty ascertaining relevance. This could create a bias towards relevance for documents with English annotation. We also realized that several judges were not correctly taking into account modality information given in the queries. For this reason we manually reviewed queries and selected some topics for rejudging. This led to results in these proceedings that are slightly different from the original working notes results. Techniques using modality detection generally performed slightly better with the revised relevance judgments. As we discovered an error in using trecval, that does not take into account rank information but only the similarity score, we also calculated a new MAP for all runs taking into account only the rank information. This is the same for many runs but a few runs become significantly better.

2.6 Submissions and Techniques

This section summarizes the main techniques used by the participants for retrieval and the sort of runs that they submitted. We had for the first time several problems with the submissions although we sent out a script to check runs for correctness before submission. In 2006, this script was part of the submission web site, but performance problems had us change this setup.

CINDI. The *CINDI* group submitted a total of 4 valid runs, two feedback runs and two automatic runs, each time one with mixed media and a purely visual run. Text retrieval uses a simple tf/idf weighting model and uses English, only. For visual retrieval a fusion model of a variety of features and image representations is used. The mixed media run simply combines the outcomes in a linear way.

DEU. *Dokuz Eylul University* submitted 5 runs, 4 visual and one textual run. The text runs is a simple bag of words approach and for visual retrieval several strategies were used containing color layout, color structure, dominant color and an edge histogram. Each run contained only one single technique.

IPAL. *IPAL* submitted 6 runs, all of them text retrieval runs. After having had the best performance for two years, the results are now only in the middle of the performance scale.

IRIT. The *IRIT* group submitted a single valid text retrieval run.

MedGIFT. The *MedGIFT* group submitted a total of 13 runs. For visual retrieval the GIFT (GNU Image Finding Tool) was used to create a baseline run as this system had been used in the same configuration since the beginning of ImageCLEF. Multilingual text retrieval was performed with EasyIR and a mapping of the text in the three languages towards MeSH (Medical Subject Headings) to search in semantic terms and avoid language problems.

MIRACLE. *MIRACLE* submitted 36 runs in total and thus most runs of all groups. The text retrieval runs were among the best, whereas visual retrieval was in the midfield. The combined runs were worse than text alone and also only in the midfield.

LIG. *MRIM-LIG* submitted 6 runs, all of them textual runs. Besides the best textual results, this was also the best overall result in 2007.

OHSU. *OHSU* submitted 10 textual and mixed runs, using Fire as a visual system. Their mixed runs had good performance as well as best early precision. Their modality detection run was the best performing mixed run.

RWTH. The human language technology and pattern recognition group from the RWTH Aachen University, Germany, submitted 10 runs using the FIRE system. The runs are based on a wide variety of 8 visual descriptors including image thumbnails, patch histograms, and various texture features. For the runs using text, a text retrieval system is used in the same way as in the last years. The weights for features are trained with a maximum entropy training method using the qrels of the 2005 and 2006 queries.

SINAI. The *SINAI* group submitted 30 runs in total, all of them textual or mixed. For text retrieval, the terms of the query are mapped onto MeSH, and then, the query is expanded with these MeSH terms.

SUNY. *SUNY* submitted 7 runs, all of which are mixed runs using Fire as visual system. One of the runs is among the best mixed runs.

UNAL. The *UNAL* group submitted 8 visual runs. The runs use a single visual feature and range towards the lower end of the performance spectrum.

MIXED. The combination of runs from *RWTH*, *OHSU*, *MedGIFT* resulted in 13 submissions, all of which were automatic and all used visual and textual information. These runs obtained a significantly better result when taking into account rank information for treceval.

2.7 Results

For the first time in 2007, the best overall official system used only text for the retrieval. Up until now the best systems always used a mix of visual and textual information. Nothing can really be said on the outcome of manual and relevance feedback submissions as there were too few submitted runs.

It became clear that most research groups participating had a single specialty, usually either visual or textual retrieval. By supplying visual and textual results as example, we gave groups the possibility to work on multi-modal retrieval as well.

Automatic Retrieval. As always, the majority of results were automatic and without any interaction. There were 146 runs in this category, with 27 visual runs, 80 mixed runs and 39 textual submissions, making automatic mixed media runs the most popular category. The results shown in the following tables are averaged over all 30 topics.

Visual Retrieval. Purely visual retrieval was performed in 27 runs and by six groups. Results from GIFT and FIRE (Flexible Image Retrieval Engine) were made available for research groups not having access to a visual retrieval engine. New MAP is the MAP calculated when taking into account rank information with treceval.

To make the tables shorter and to not bias results shown towards groups with many submissions, only the best two and the worst two runs of every group are shown in the tables. Table 2 shows the results for the visual runs. Most runs had an extremely low MAP (<3% MAP), which had been the case during the previous years as well. The overall results were lower than in preceding years, indicating that tasks might have become harder. On the other hand, two runs had good results and rivaled, at least for early precision, the best textual results. These two runs used data from 2005 and 2006 that was somewhat similar to the tasks in 2007 to train the system for optimal feature selection. This showed that an optimized feature weighting may result in a large improvement!

Textual Retrieval. A total of 39 submissions were purely textual and came from nine research groups. Table 3 shows the best and worst two results of every group for purely textual retrieval. The best overall runs were from LIG and were purely textual, which happened for the first time in ImageCLEF. LIG participated in

Table 2. Automatic runs using visual information (best/worst two of every group)

Run	Relevant	MAP	new MAP	bpref	P5	P10	P30
RWTH-FIRE-ME-NT-tr0506	1376	0.2427	0.2426	0.283	0.48	0.45	0.3756
RWTH-FIRE-ME-NT-tr06	1368	0.23	0.2300	0.2696	0.48	0.4467	0.3722
CINDI_IMG_FUSION	567	0.0355	0.0354	0.0751	0.1533	0.1233	0.1122
RWTH-FIRE-NT-emp	506	0.0264	0.0264	0.056	0.0933	0.0933	0.0744
RWTH-FIRE-NT-emp2	474	0.0255	0.0255	0.0535	0.1067	0.0933	0.0656
miracleVisG	496	0.0182	0.0182	0.0448	0.0933	0.08	0.0767
miracleVisGFANDmm	156	0.01	0.01	0.0221	0.0667	0.0667	0.05
miracleVisGFANDavg	156	0.0085	0.0085	0.0185	0.0467	0.0467	0.0556
miracleVisGFANDmin	156	0.0079	0.0079	0.0184	0.04	0.0367	0.0478
UNALCO-nni_Sobel	433	0.0072	0.0076	0.0668	0.02	0.02	0.0133
UNALCO-nni_FeatComb	531	0.0066	0.0205	0.0825	0.0133	0.02	0.0122
DEU_CS-DEU_R2	239	0.0062	0.0111	0.0433	0.0133	0.0067	0.0022
UNALCO-svmRBF_RGBHis	329	0.0048	0.0135	0.0481	0.0133	0.0133	0.0089
UNALCO-svmRBF_Tamura	341	0.0046	0.0055	0.0536	0.0133	0.0067	0.01
GE_4.8	245	0.0035	0.0035	0.0241	0.04	0.0333	0.0233
GE-GE_GIFT4	244	0.0035	0.0035	0.024	0.04	0.0333	0.0233
GE-GE_GIFT8	245	0.0035	0.0035	0.024	0.04	0.0333	0.0233
DEU_CS-DEU_R4	199	0.0017	0.0035	0.04	0.0067	0.0033	0.0056
DEU_CS-DEU_R3	216	0.0016	0.0079	0.0442	0.0067	0.01	0.0056
DEU_CS-DEU_R5	195	0.0013	0.0038	0.0351	0	0	0.0078

ImageCLEF this year for the first time. Early precision (P5) was similar to the best purely visual runs and the best mixed runs had a very high early precision. The highest P10 was a mixed system where the MAP was situated lower. Despite its name, MAP is more of a recall-oriented measure. Re-scoring of the results with treceval basing the order of documents on the rank results in a few runs becoming significantly better but does not change many of the other runs.

Mixed Retrieval. Mixed automatic retrieval had the highest number of submissions of all categories. There were 80 runs submitted by 8 participating groups.

Table 4 summarizes the best two and the worst two mixed runs of every group. For some groups the results for mixed runs were better than the best text runs but for others this was not the case. This underlines the fact that combinations between visual and textual features have to be done with care. Another interesting fact is that some systems with only a mediocre MAP performed extremely well with respect to early precision. All early precision values (P5, P10, P30) had their best results with mixed submissions.

Another interesting fact could be observed after correctly rescoring the results as the best mixed run is in this case much better than the best textual run. All combination runs of gift, fire, and ohsu obtain extremely much better results bringing them up the performing runs.

Table 3. Automatic runs using only text (best and worst two of every group)

Run	Relevant	MAP	new MAP	bpref	P5	P10	P30
LIG-MRIM-LIG_MU_A	1904	0.3538	0.3533	0.3954	0.42	0.43	0.3844
LIG-MRIM-LIG_GM_A	1898	0.3517	0.3513	0.395	0.42	0.4233	0.3922
miracleTxtENN	1842	0.3385	0.3427	0.406	0.4933	0.4567	0.3578
LIG-MRIM-LIG_GM_L	1909	0.3345	0.3338	0.3855	0.4467	0.4433	0.3856
ohsu_text_e4_out_rev1	1459	0.3317	0.3467	0.3957	0.46	0.4733	0.3956
LIG-MRIM-LIG_MU_L	1912	0.3269	0.3263	0.3802	0.44	0.4333	0.3656
OHSU-OHSU_txt_exp2	1162	0.3192	0.3339	0.3688	0.46	0.4733	0.3956
SinaiC100T100	1985	0.2944	0.3052	0.3505	0.3933	0.4367	0.3967
UB-NLM-UBTextBL1	1825	0.2897	0.2897	0.3279	0.3867	0.41	0.3678
SinaiC040T100	1937	0.2838	0.2978	0.3269	0.4067	0.4533	0.4033
IPAL1.TXT_BAY_ISA0.2	1515	0.2784	0.2781	0.323	0.42	0.39	0.31
IPAL1.TXT_BAY_ISA0.1	1517	0.2783	0.278	0.3233	0.4133	0.39	0.3122
OHSU-as_out_1000rev1_c	1871	0.2754	0.2799	0.3346	0.44	0.4367	0.36
OHSU-oshu_as_is_1000	1871	0.2754	0.2816	0.3345	0.44	0.4367	0.36
IPAL.TXT_BAY_ALLREL2	1520	0.275	0.2746	0.3215	0.4067	0.3767	0.3122
IPAL4.TXT_BAY_ISA0.4	1468	0.2711	0.2708	0.3218	0.3933	0.3867	0.3078
SinaiC030T100	1910	0.271	0.2748	0.3126	0.42	0.41	0.3822
miracleTxtXN	1784	0.2647	0.2659	0.3711	0.3267	0.3367	0.3167
UB-NLM-UBTextBL2	1666	0.2436	0.2437	0.2921	0.3133	0.3033	0.2811
GE_EN	1839	0.2369	0.2373	0.2867	0.2867	0.3333	0.2678
SinaiC020T100	1589	0.2356	0.2366	0.2665	0.34	0.3467	0.3422
GE_MIX	1806	0.2186	0.2192	0.2566	0.3133	0.2967	0.2622
DEU_CS-DEU_R1	727	0.1611	0.1618	0.1876	0.3067	0.32	0.3033
GE_DE	1166	0.1433	0.1441	0.209	0.2267	0.2	0.15
UB-NLM-UBTextFR	1248	0.1414	0.1413	0.2931	0.2	0.1933	0.1533
GE_FR	1139	0.115	0.115	0.1503	0.1	0.1267	0.1289
miracleTxtFRT	906	0.0863	0.085	0.1195	0.1733	0.1733	0.15
miracleTxtFRN	815	0.0846	0.0822	0.1221	0.26	0.18	0.1367
IRIT_RunMed1	1163	0.0486	0.1201	0.1682	0.0533	0.05	0.0756

2.8 Manual and Interactive Retrieval

Only three runs were in the manual or interactive sections, making any real comparison impossible. Table 5 lists these runs and their performance. Although information retrieval with relevance feedback or manual query modifications are thought to be a very important area to improve performance, research groups in ImageCLEF 2007 did not make use of it.

2.9 Conclusions

Visual retrieval without learning had very low results for MAP and even for early precision (although with a smaller difference from text retrieval). Visual topics perform well using visual techniques. Extensive learning of feature selection and weighting can have enormous gain in performance as shown by FIRE.

Table 4. Automatic runs using mixed information (best and worst two of every group)

Run	Relevant	MAP	new MAP	bpref	P5	P10	P30
ohsu_m2_rev1_c	1778	0.3415	0.4084	0.4099	0.4467	0.4333	0.37
SinaiC100T80	1976	0.2999	0.3026	0.3425	0.4	0.4567	0.4067
RWTH-FIRE-ME-tr0506	1566	0.2962	0.2962	0.3414	0.4733	0.4667	0.3978
RWTH-FIRE-ME-tr06	1566	0.296	0.296	0.3407	0.4933	0.47	0.3978
UB-NLM-UBTL3	1833	0.2938	0.2938	0.3306	0.3867	0.4167	0.3689
UB-NLM-UBTL1	1831	0.293	0.2928	0.335	0.3867	0.4	0.3867
SinaiC040T80	1948	0.2914	0.2949	0.3236	0.4267	0.4667	0.4133
UB-NLM-UBmixedMulti2	1666	0.2537	0.2537	0.3011	0.3467	0.3167	0.29
miracleMixGENTRIGHTmin	1608	0.248	0.2439	0.2936	0.3667	0.3533	0.3011
RWTH-FIRE-emp2	1520	0.2302	0.2302	0.2803	0.3867	0.4	0.3689
RWTH-FIRE-emp	1521	0.2261	0.2261	0.2758	0.38	0.4	0.3711
miracleMixGENTRIGHTmax	1648	0.2225	0.2259	0.2687	0.3067	0.32	0.2856
GE_VT1_4	1806	0.2195	0.2199	0.2567	0.32	0.3033	0.2622
GE_VT1_8	1806	0.2195	0.2204	0.2566	0.32	0.3033	0.2622
OHSU-ohsu_m1	509	0.2167	0.2374	0.2405	0.3867	0.3933	0.3567
CINDI.TXT.IMAGE.LINEAR	944	0.1906	0.1914	0.2425	0.34	0.3133	0.2822
SinaiC060T50	1863	0.1874	0.1882	0.2245	0.4	0.3767	0.2789
GE_VT10_4	1192	0.1828	0.1829	0.2141	0.3	0.31	0.2633
GE_VT10_8	1196	0.1828	0.1839	0.214	0.3	0.31	0.2633
SinaiC020T50.clef	1544	0.1727	0.1726	0.1967	0.3133	0.3267	0.2744
UB-NLM-UBmixedFR	997	0.1364	0.1363	0.2168	0.2133	0.2	0.1789
ohsu_comb3_ef_wt1_rev1_c	903	0.1113	0.1144	0.1525	0.2533	0.2433	0.1522
ohsu_fire_ef_wt2_rev1_c	519	0.0577	0.0608	0.0888	0.16	0.16	0.1122
3fire-7ohsu	1887	0.0303	0.2355	0.1115	0.0067	0.01	0.0067
5fire-5ohsu	1892	0.0291	0.2871	0.1012	0.0067	0.0067	0.0078
5gift-5ohsu	1317	0.0153	0.1867	0.1151	0	0.0033	0.0022
7gift-3ohsu	1319	0.0148	0.2652	0.1033	0	0.0033	0.0022
miracleGFANDminLEFTmm	156	0.0097	0.0097	0.0197	0.0533	0.0533	0.0544
miracleGFANDminLEFTmax	156	0.0079	0.0079	0.0184	0.04	0.0367	0.0478

Table 5. The only three runs not using automatic retrieval

Run	Rel.	MAP	new	bpref	P10	P30	media	interaction
CINDI_IMG_FUSION_RF	610	0.04	0.04	0.09	0.15	0.119	visual	feedback
CINDI.TXT_IMG_RF.LIN	773	0.12	0.12	0.19	0.36	0.251	mixed	feedback
OHSU-oshu_man2	1795	0.35	0.36	0.40	0.443	0.349	textual	manual

Purely textual runs had the best overall results for the first time and text retrieval was shown to work well for most topics. Mixed-media runs were the most popular category and are often better in performance than text or visual features alone. When correctly scoring all runs the best performance was actually in this category. Still, in many cases the mixed media runs did not perform as

well as text alone, showing that care needs to be taken to combine media. These runs do have the best performance for all early precision values.

Interactive and manual queries were almost absent from the evaluation and this remains an important problem. ImageCLEFmed has to put these domains more into the focus of the researchers although this requires more resources to perform the evaluation. System-oriented evaluation is an important part but only interactive retrieval can show how well a system can really help the users.

With respect to performance measures, there was less correlation between the measures than in previous years. The runs with the best early precision (P10) were not as good in MAP to the best overall systems. This needs to be investigated as MAP is indeed a good indicator for overall system performance but early precision might be much more what real users are looking for.

3 The Medical Automatic Annotation Task

Over the last two years, automatic medical image annotation has been evolved from a simple classification task with about 60 classes to a task with about 120 classes. From the very start however, it was clear that the number of classes cannot be scaled indefinitely, and that the number of classes that are desirable to be recognised in medical applications is far too big to assemble sufficient training data to create suitable classifiers. To address this issue, a hierarchical class structure such as the IRMA code [14] can be a solution which allows to create a set of classifiers for subproblems. The classes in the last years were based on the IRMA code where created by grouping similar codes in one class. This year, the task has changed and the objective is to predict complete IRMA codes instead of simple classes.

This year's medical automatic annotation task builds on top of last year: 1,000 new images were collected and are used as test data, the training and the test data of last year was used as training and development data respectively.

3.1 Database and Task Description

The complete database consists of 12'000 fully classified medical radiographs taken randomly from medical routine at the RWTH Aachen University Hospital. 10'000 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 11'000 images (10'000 training + 1'000 validation) as training data.

Each of the 12'000 images is annotated with its complete IRMA code (see Sec. 3.1). In total, 116 different IRMA codes occur in the database, the codes are not uniformly distributed, but some codes have a significant larger share among the

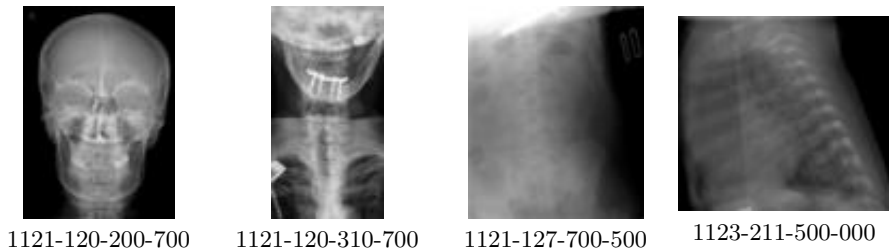


Fig. 3. Example images from the medical annotation task with full IRMA-code. The textual representation of the IRMA codes is (from left to right):

T: x-ray, plain radiography, analog, overview image; D: coronal, anteroposterior (AP, coronal), unspecified; A: cranium, unspecified, unspecified; B: musculoskeletal system, unspecified, unspecified.

T: x-ray, plain radiography, analog, overview image; D: coronal, anteroposterior (AP, coronal), unspecified; A: spine, cervical spine, unspecified; B: musculoskeletal system, unspecified, unspecified.

T: x-ray, plain radiography, analog, overview image; D: coronal, anteroposterior (AP, coronal), supine; A: abdomen, unspecified, unspecified; B: uropoietic system, unspecified, unspecified.

T: x-ray, plain radiography, analog, high beam energy; D: sagittal, lateral, right-left, inspiration; A: chest, unspecified, unspecified; B: unspecified, unspecified, unspecified.

data than others. The least frequent codes however, are represented at least 10 times in the training data to allow for learning suitable models.

Example images from the database together with textual labels and their complete code are given in Figure 3.

IRMA Code. Existing medical terminologies such as the MeSH thesaurus are poly-hierarchical, i.e., a code entity can be reached over several paths. However, in the field of content-based image retrieval, we frequently find class-subclass relations. The mono-hierarchical multi-axial IRMA code strictly relies on such part-of hierarchies and, therefore, avoids ambiguities in textual classification [14]. In particular, the IRMA code is composed from four axes having three to four positions, each in $\{0, \dots, 9, a, \dots, z\}$, where "0" denotes "not further specified". More precisely,

- the technical code (T) describes the imaging modality;
- the directional code (D) models body orientations;
- the anatomical code (A) refers to the body region examined; and
- the biological code (B) describes the biological system examined.

This results in a string of 13 characters (IRMA: TTTT – DDD – AAA – BBB). For instance, the body region (anatomy, three code positions) is defined as follows:

AAA

000 not further specified

...

400 upper extrimity (arm)

410 upper extrimity (arm); hand

411 upper extrimity (arm); hand; finger

412 upper extrimity (arm); hand; middle hand

413 upper extrimity (arm); hand; carpal bones

420 upper extrimity (arm); radio carpal joint

430 upper extrimity (arm); forearm

431 upper extrimity (arm); forearm; distal forearm

432 upper extrimity (arm); forearm; proximal forearm

440 upper extrimity (arm); ellbow

...

The IRMA code can be easily extended by introducing characters in a certain code position, e.g., if new imaging modalities are introduced. Based on the hierarchy, the more code position differ from "0", the more detailed is the description.

Hierarchical Classification. To define a evaluation scheme for hierarchical classification, we can consider the 4 axes to be uncorrelated. Hence, we assume the axes independently and just sum up the errors for each axis independently.

Hierarchical classification is a well-known topic in different field. For example the classification of documents often is done using an ontology-based class hierarchy [15] and in information extraction similar techniques are applied [16]. In our case, however we developed a novel evaluation scheme to account for the particularities of the IRMA code which considers errors that are made early in a hierarchy to be worse than errors that are made at a fine level, and it is explicitly possible to predict a code partially, i.e. to predict a code up to a certain position and put wild-cards for the remaining positions, which is penalised but only with half the penalty a misclassification is penalised.

Our evaluation scheme is described in the following, where we only consider one axis. The same scheme is applied to each axis individually.

Let $l_1^I = l_1, l_2, \dots, l_i, \dots, l_I$ be the *correct* code (for one axis) of an image, i.e. if a classifier predicts this code for an image, the classification is perfect. Further, let $\hat{l}_1^I = \hat{l}_1, \hat{l}_2, \dots, \hat{l}_i, \dots, \hat{l}_I$ be the *predicted* code (for one axis) of an image.

The correct code is specified completely: l_i is specified for each position. The classifiers however, are allowed to specify codes only up to a certain level, and predict "*don't know*" (encoded by $*$) for the remaining levels of this axis.

Given an incorrect classification at position $\hat{l}_i$ we consider all succeeding decisions to be wrong and given a not specified position, we consider all succeeding decisions to be not specified.

We want to penalise wrong decisions that are easy (fewer possible choices at that node) over wrong decisions that are difficult (many possible choices at that node), we can say, a decision at position l_i is correct by chance with a probability

of $\frac{1}{b_i}$ if b_i is the number of possible labels for position i . This assumes equal priors for each class at each position.

Furthermore, we want to penalise wrong decisions at an early stage in the code (higher up in the hierarchy) over wrong decisions at a later stage in the code (lower down on the hierarchy) (i.e. l_i is more important than l_{i+1}).

Assembling the ideas from above straight forwardly leads to the following equation:

$$\sum_{i=1}^I \underbrace{\frac{1}{b_i}}_{(a)} \underbrace{\frac{1}{i}}_{(b)} \underbrace{\delta(l_i, \hat{l}_i)}_{(c)}$$

with

$$\delta(l_i, \hat{l}_i) = \begin{cases} 0 & \text{if } l_j = \hat{l}_j \quad \forall j \leq i \\ 0.5 & \text{if } l_j = * \quad \exists j \leq i \\ 1 & \text{if } l_j \neq \hat{l}_j \quad \exists j \leq i \end{cases}$$

where the parts of the equation account for

- (a) accounts for difficulty of the decision at position i (branching factor)
- (b) accounts for the level in the hierarchy (position in the string)
- (c) correct/not specified/wrong, respectively.

In addition, for every code, the maximal possible error is calculated and the errors are normed such that a fully incorrect decision (i.e. all positions wrong) gets an error count of 1.0 and an image classified correctly in all positions has an error of 0.0.

Table 6 shows examples for a correct code with different predicted codes. Predicting the completely correct code leads to an error measure of 0.0, predicting all positions incorrectly leads to an error measure of 1.0. The examples in Table 6 demonstrate that a classification error in a position at the back of the code results in a lower error measure than a position in one of the first positions. The last column of the table show the effect of the branching factor b . In this column we assumed the branching factor of the code is $b = 2$ in each node of the hierarchy. It can be observed that the errors for the later positions have more weight compared to the real errors in the real hierarchy.

Table 6. Example scores for hierarchical classification, based on the correct code IRMA TTTT = 318a and assuming the branching factor would be 2 in each node of the hie

classified	error measure	error measure (b=2)
318a	0.000	0.000
318*	0.024	0.060
3187	0.049	0.120
31*a	0.082	0.140
31**	0.082	0.140
3177	0.165	0.280
3***	0.343	0.260
32**	0.687	0.520
1000	1.000	1.000

3.2 Participating Groups and Methods

In the medical automatic annotation task, 29 groups registered of which 10 groups participated, submitting a total of 68 runs. The group with the highest number of submissions had 30 runs in total.

In the following, groups are listed alphabetically and their methods are described shortly.

BIOMOD: University of Liege, Belgium. The Bioinformatics and Modelling group from the University Liege in Belgium submitted four runs. The approach is based on an object recognition framework using extremely randomised trees and randomly extracted sub-windows [17]. All runs use the same technique but differ how the code is assembled.

BLOOM: IDIAP, Switzerland. The Blanceflor-om2-toMed group from IDIAP in Martigny, Switzerland submitted 7 runs. All runs use support vector machines (either in one-against-one or one-against-the-rest manner). Features used are downsampled versions of the images, SIFT features extracted from sub-images, and combinations of these [18].

Geneva: medGIFT Group, Switzerland. The medGIFT group from Geneva, Switzerland submitted 3 runs, each of the runs uses the GIFT image retrieval system. The runs differ in the way, the IRMA-codes of the top-ranked images are combined [19].

CYU: Information Management AI lab, Taiwan. The Information Management AI lab from the Ching Yun University of Jung-Li, Taiwan submitted one run using a nearest neighbour classifier using different global and local image features which are particularly robust with respect to lighting changes.

MIRACLE: Madrid, Spain. The Miracle group from Madrid, Spain submitted 30 runs. The classification was done using a 10-nearest neighbour classifier and the features used are gray-value histograms, Tamura texture features, global texture features, and Gabor features, which were extracted using FIRE. The runs differ which features were used and how the prediction of the code was done.

Oregon Health State University, Portland, OR, USA. The Department of Medical Informatics and Clinical Epidemiology of the Oregon Health and Science University in Portland, Oregon submitted two runs using neural networks and GIST descriptors. One of the runs uses a support vector machine as a second level classifier to help discriminating the two most difficult classes.

RWTHi6: RWTH Aachen University, Aachen, Germany. The Human Language Technology and Pattern Recognition group of the RWTH Aachen University in Aachen, Germany submitted 6 runs, all are based on sparse histograms of image patches which were obtained by extracting patches at each position in the image [20]. One run is a combination of 4 normal runs, and one run does the classification axis-wise.

IRMA: RWTH Aachen University, Medical Informatics, Aachen, Germany. The IRMA group from the RWTH Aachen University Hospital in

Aachen, Germany submitted three baseline runs using weighted combinations of nearest neighbour classifiers using texture histograms, image cross correlations, and the image deformation model. The parameters used are exactly the same as used in previous years. The runs differ in the way in which the codes of the five nearest neighbours are used to assemble the final predicted code.

UFR: University of Freiburg, Computer Science Dep., Freiburg, Germany. The Pattern Recognition and Image Processing group from the University Freiburg, Germany, submitted four runs using relational features calculated around interest points which are later combined to form cluster cooccurrence matrices [21]. Three different classification methods were used.

UNIBAS: University of Basel, Switzerland. The Databases and Information Systems group from the University Basel, Switzerland submitted 14 runs using a pseudo two-dimensional hidden Markov model to model image deformation in the images which were scaled down keeping the aspect ratio such that the longer side has a length of 32 pixels [23].

3.3 Results

An overview of the results of the evaluation is given in Table 7. For each group, the number of submissions, the best and the worst rank, the minimal and the maximal score, the mean and the median score, the best and the worst error rate, the mean and the median error rate are given.

The method which had the best result last year is now at rank 8, which gives an impression how much improvement in this field was achieved over the last year.

Looking at the results for individual images, we noted, that only one image was classified correctly by all submitted runs (top left image in Fig. 3). No image was misclassified by all runs.

3.4 Discussion

Analysing the results, it can be observed that the top-performing runs do not consider the hierarchical structure of the given task, but rather use each individual code as one class and train a 116 classes classifier. This approach seems to work better given the currently limited amount of codes, but obviously would not scale up infinitely and would probably lead to a very high demand for appropriate training data if a much larger amount of classes is to be distinguished. The best run using the code is on rank 6, builds on top of the other runs from the same group and uses the hierarchy only in a second stage to combine the four runs.

Furthermore, it can be seen that a method that is applied once accounting for the hierarchy/axis structure of the code and once using the straight forward classification into 116 classes approach, the one which does not know about the hierarchy clearly outperforms the other one (runs on ranks 11 and 13/7 and 14,16).

Table 7. Results of the evaluation by participating group. For each group, the number of submitted runs, the rank of the best and worst run, and the minimum, maximum, mean, and medium error count and error rate are given.

group	# sub	rank		score				ER			
		min	max	min	max	mean	median	min	max	mean	median
BIOMOD	4	30	35	73.82	95.25	80.90	77.26	22.90	36.00	29.28	29.10
BLOOM	7	1	29	26.85	72.41	40.44	29.46	10.30	20.80	13.77	11.50
Geneva	3	63	65	375.72	391.02	385.68	390.29	99.00	99.70	99.33	99.30
CYU	1	33	33	79.30	79.30	79.30	79.30	25.30	25.30	25.30	25.30
MIRACLE	30	36	68	158.82	505.62	237.42	196.18	49.30	89.00	62.09	55.50
OHSU	2	26	27	67.81	67.98	67.89	67.89	22.70	22.70	22.70	22.70
RWTHi6	6	6	13	30.93	44.56	35.16	33.88	11.90	17.80	13.38	12.55
IRMA	3	17	34	51.34	80.47	61.45	52.54	18.00	45.90	27.97	20.00
UFR	5	7	16	31.44	48.41	41.29	45.48	12.10	17.90	15.36	16.80
UNIBAS	7	19	25	58.15	65.09	61.64	61.41	20.20	23.20	22.26	22.50

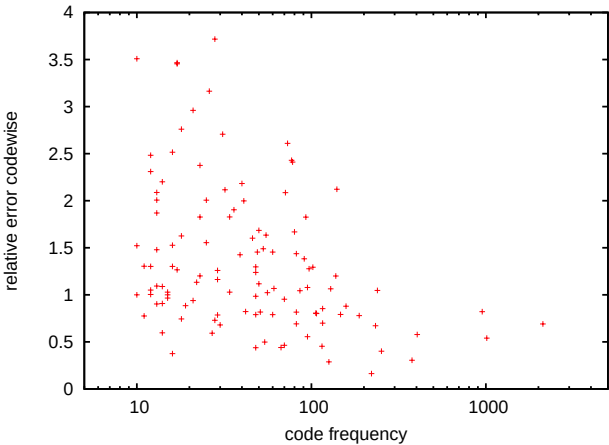


Fig. 4. Code-wise relative error as a function of the frequency of this code in the training data

Another clear observation is that methods using local image descriptors outperform methods using global image descriptors. In particular, the top 16 runs are all using either local image features alone or local image features in combination with a global descriptor.

It is also observed that images where a large amount of training data is available are more far more likely to be classified correctly.

Considering the ranking according to the applied hierarchical measure and the ranking according to the error rate it can clearly be seen that there are hardly any differences. Most of the differences are clearly due to use of the code (mostly inserting of wildcard characters) which can lead to an improvement for the hierarchical evaluation scheme, but will always lead to a deterioration of the error rate.

3.5 Conclusion

The success of the medical automatic annotation task could be continued, the number of participants is pretty constant, but a clear performance improvement of the best method could be observed. Although only few groups actively tried to exploit the hierarchical class structure many of the participants told us that they consider this an important research topic and that a further investigation is desired.

Our goal for future tasks is to motivate more groups to participate and to increase the database size such that it is necessary to use the hierarchical class structure actively.

4 Overall Conclusions

The two medical tasks of ImageCLEF again attracted a very large number of registrations and participation. This underlines the importance of such evaluation campaigns giving researchers the opportunity to evaluate their systems without the tedious task of creating databases and topics. In domains such as medical retrieval this is particularly important as data access is often difficult.

In the medical retrieval task, visual retrieval without any learning only obtained good results for a small subset of topics. With learning this can change strongly and deliver even for purely visual retrieval very good results. Mixed-media retrieval was the most popular category and results were often better for mixed-media than textual runs of the same groups. This shows that mixed-media retrieval requires much work and more needs to be learned on such combinations. The best systems concerning early precision were mixed media runs. Interactive retrieval and manual query modification were only used in 3 out of the 149 submitted runs. This shows that research groups prefer submitting automatic runs, although interactive retrieval is important and still must be addressed by researchers.

For the annotation task, it was observed that techniques that rely heavily on recent developments in machine learning and build on modern image descriptors clearly outperform other methods. The class hierarchy that was provided could only lead to improvements for a few groups. Overall, the runs that use the class hierarchy perform worse than those, which consider every code as a unique class giving the impression that for the current number of 116 unique codes the training data is sufficient to train a joint classifier.

Acknowledgements

We thank CLEF for supporting ImageCLEF. We also thank all organizations who provided images and annotations for this year's task, including myPACS.net (Rex Jakobovits) and the OHSU CORI project (Judith Logan).

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 Sixth Framework Program with the SemanticMining project (IST NoE 507505) and the 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., Jones, G.J.F., Gonzalo, J., Kluck, M., Magnini, B. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 243–251. Springer, Heidelberg (2005)
2. Clough, P., Müller, H., Sanderson, M.: The CLEF 2004 cross language image retrieval track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
3. Savoy, J.: Report on CLEF-2001 experiments. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 27–43. Springer, Heidelberg (2002)
4. 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 (September 2006)
5. Smeaton, A.F., Over, P., Kraaij, W.: TRECVID: Evaluating the effectiveness of information retrieval tasks on digital video. In: Proceedings of the international ACM conference on Multimedia 2004 (ACM MM 2004), New York City, NY, USA, October 2004, pp. 652–655 (2004)
6. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
7. Deselaers, T., Hanbury, A., et al.: Overview of the ImageCLEF 2007 object retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
8. 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 (September/October 2006)
9. Müller, H., Rosset, A., Vallée, J.-P., Terrier, F., Geissbuhler, A.: A reference data set for the evaluation of medical image retrieval systems. Computerized Medical Imaging and Graphics 28, 295–305 (2004)
10. 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(2), 1–6 (2004)
11. Candler, C.S., Uijtdehaage, S.H., Dennis, S.E.: Introducing HEAL: The health education assets library. Academic Medicine 78(3), 249–253 (2003)
12. Wallis, J.W., Miller, M.M., Miller, T.R., Vreeland, T.H.: An internet-based nuclear medicine teaching file. Journal of Nuclear Medicine 36(8), 1520–1527 (1995)
13. Glatz-Krieger, K., Glatz, D., Gysel, M., Dittler, M., Mihatsch, M.J.: Webbasierte Lernwerkzeuge für die Pathologie – web-based learning tools for pathology. Pathologie 24, 394–399 (2003)

14. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: SPIE 2003, vol. 5033, pp. 440–451 (2003)
15. Sun, A., Lim, E.P.: Hierarchical text classification and evaluation. In: IEEE International Conference on Data Mining (ICDM 2001), San Jose, CA, USA, November 2001, pp. 521–528 (2001)
16. Maynard, D., Peters, W., Li, Y.: Metrics for evaluation of ontology-based information extraction. In: Evaluation of Ontologies for the Web (EON 2006), Edinburgh, UK (2006)
17. Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Random subwindows for robust image classification. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005), June 2005, vol. 1, pp. 34–40. IEEE, Los Alamitos (2005)
18. Tommasi, T., Orabona, F., Caputo, B.: CLEF2007 Image Annotation Task: an SVM-based Cue Integration Approach. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
19. Zhou, X., Gobeill, J., Ruch, P., Müller, H.: University and Hospitals of Geneva at ImageCLEF 2007. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
20. Deselaers, T., Hegerath, A., Keysers, D., Ney, H.: Sparse patch-histograms for object classification in cluttered images. In: Franke, K., Müller, K.-R., Nikolay, B., Schäfer, R. (eds.) DAGM 2006. LNCS, vol. 4174, pp. 202–211. Springer, Heidelberg (2006)
21. 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, Santa Barbara, CA, USA (2006)
22. Setia, L., Burkhardt, H.: Learning taxonomies in large image databases. In: ACM SIGIR Workshop on Multimedia Information Retrieval, Amsterdam, Holland (2007)
23. Springmann, M., Schuldt, H.: Speeding up idm without degradation of retrieval quality. In: Nardi, A., Peters, C. (eds.) Working Notes of the CLEF Workshop 2007 (2007)

FIRE in ImageCLEF 2007: Support Vector Machines and Logistic Models to Fuse Image Descriptors for Photo Retrieval

Tobias Gass, Tobias Weyand, Thomas Deselaers, and Hermann Ney

Human Language Technology and Pattern Recognition Group,
RWTH Aachen University, Aachen, Germany
{last name}@i6.informatik.rwth-aachen.de

Abstract. Submissions to the photographic retrieval task of the ImageCLEF 2007 evaluation and improvements of our methods that were tested and evaluated after the official benchmark. We use our image retrieval system FIRE to combine a set of different image descriptors. The most important step in combining descriptors is to find a suitable weighting. Here, we evaluate empirically tuned linear combinations, a trained logistic regression model, and support vector machines to fuse the different descriptors. Additionally, clustered SIFT histograms are evaluated for the given task and show very good results – both, alone and in combination with other features. A clear improvement over our evaluation performance is shown consistently over different combination schemes and feature sets.

Keywords: content-based image retrieval, feature combination, SIFT features.

1 Introduction

ImageCLEF¹ is an evaluation event for textual (mono- and multilingual) and content-based retrieval of images. Evaluation campaigns are an important factor to foster progress in research and therefore we participated for the third time using our content-based image retrieval system FIRE.

Although the machine learning community has produced a large amount of strong classification techniques, the image retrieval community has so far only employed k-nearest neighbor approaches or techniques derived from (textual) information retrieval. Thus, only few systems build on top of state-of-the-art machine learning and image representation techniques. In this paper we evaluate histograms of SIFT features, a common image descriptor for object recognition, which is so far seldomly used for general photographic image retrieval, and compare different strategies to fuse image descriptors. The considered strategies are maximum entropy-based feature combination as presented last year [1] and support vector machines.

¹ <http://www.imageclef.org>

Support vector machines have so far been used in image retrieval by [2] for relevance feedback and as a feature combination strategy. A similar approach is presented in [3].

2 ImageCLEF 2007 Photographic Retrieval Task

The database used in the photographic retrieval task [4] was the IAPR TC-12 photographic collection [5] consisting of 20,000 natural still images annotated in three languages. 60 queries were given, each consisting of a short textual description and three sample images. The queries posed in 2007 are very similar to the 2006 queries. This allows the 2006 queries to be used in combination with the relevance judgements to train the log-linear models and SVMs for feature combination.

3 Features

In this section, we present the image descriptors we used in our experiments.

We briefly outline the features we used for our runs in the ImageCLEF 2007 evaluation. Additionally, we present *Clustered SIFT Histograms*, a variant of clustered histograms of local features, which will result in clear improvements over our official evaluation results.

Colour Histograms. Colour histograms are among the most basic approaches and widely used in image retrieval [6]. The colour space is partitioned and for each partition the pixels with a colour within its range are counted, resulting in a representation of the relative frequencies of the occurring colours. In accordance to [7], we use Jeffrey divergence to compare histogram descriptors.

Global Texture Descriptor. In [8] a texture feature is described consisting of several parts: *Fractal dimension*, *Coarseness*, *Entropy*, *spatial Gray-level difference statistics*, and the *circular Moran autocorrelation function*. From these, we obtain 43 dimensional vectors which have been successfully used in preceding ImageCLEF evaluations.

Invariant Feature Histograms. A feature is called invariant with respect to certain transformations if it does not change when these transformations are applied to the image. The transformations considered here are translation, rotation, and scaling. In this work, invariant feature histograms as presented in [9] are used.

Tamura Features. In [10] the authors propose six texture features corresponding to human visual perception: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity*, and *roughness*. In our experiments we use coarseness, contrast, and directionality to create a histogram describing the texture [8].

GIFT Colour Descriptors. In [11] the authors propose global and local colour features using a quantised HSV colour space. The global features are colour

histograms over the complete image, the local features are extracted from a uniform grid and describe the dominating colours in each grid cell.

Clustered Histograms of Local Features. Local image features offer some advantages over the mainly global image descriptors presented so far, because they allow for matching images w.r.t. the local concepts they have in common. To be able to cope with the amount of data that is to be handled when local image descriptors are extracted in great numbers, we represent the local descriptors in a histogram over a dictionary of visual words as proposed in [12]. As opposed to [12], we do not simply use image patches alone but apply the same technique to SIFT features [13], which have been shown to outperform image patches in many applications [14].

The creation of these histograms is a three step procedure:

1. local features are extracted from all training images,
2. the local features of all training images are jointly clustered using the EM algorithm for Gaussian mixtures to form 256-8000 clusters,
3. all information about each local feature 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 words from the code-book occur in the image.

These histograms are created using SIFT features and image patches. We reduce the dimensionality of the local features to 40 using PCA transformation to reduce the amount of data.

The histograms over SIFT features were not used in the official evaluation. In section 5, we show that they can lead to a performance improvement.

Sparse Patch Histograms. A computationally more efficient method for generating patch histograms was proposed in [15]. First, all patches are transformed into a lower-dimensional space using PCA. Then, a histogram grid is estimated by calculating the mean and variance of each axis of this space. Now, a patch histogram is created for each image by inserting all patches from the image into this grid. This technique allows to skip the computationally expensive step of creating a visual vocabulary by spanning the complete feature space, but does not perform as good as clustered histograms since the created histograms are sparse and the bins do not represent visual words.

4 Image Retrieval Method

Given a set of positive example images $\mathcal{Q}^+$ and a (possibly empty) set of negative example images $\mathcal{Q}^-$ a score $S(\mathcal{Q}^+, \mathcal{Q}^-, X)$ is calculated for each image X from the database:

$$S(\mathcal{Q}^+, \mathcal{Q}^-, X) = \sum_{Q \in \mathcal{Q}^+} S(Q, X) + \sum_{Q \in \mathcal{Q}^-} (1 - S(Q, X)). \quad (1)$$

where $S(Q, X)$ is the score of database image X with respect to query Q and is calculated as $S(Q, X) = e^{-\gamma W(Q, X)}$ with $\gamma = 1.0$. $W(Q, X)$ is a weighted sum of distances calculated as

$$D(Q, X) := \sum_{i=1}^I w_i \cdot d_i(Q_i, X_i). \quad (2)$$

Here, Q_i and X_i are the i th feature of the query image Q and the database image X , respectively. d_i is the corresponding distance measure and w_i is a weighting coefficient. For each d_i , $\sum_{X \in \mathcal{B}} d_i(Q_i, X_i) = 1$ is enforced by re-normalisation.

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 retrieval engine.

The selection of the weights w_i is a critical step, for which two methods have been deployed so far. It is possible to heuristically choose feature weights based on the performance of the individual features. This usually leads to superior results than an unweighted baseline. The second approach uses the maximum entropy framework to train a logistic model, which can then be used to calculate a score for a query/database image pair. Additionally, we present a novel approach to weight features using support vector machines.

4.1 Scoring of Distances Using Classifiers

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 +1) or irrelevant (denoted by -1)

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 each query image Q_n of the 2006 task the distance vector $D(Q_m, X_n) = (d_1(Q_{m1}, X_{n1}), \dots, d_I(Q_{mI}, X_{nI}))$ are calculated. This leads to N distance vectors for each of the images Q_m , where N denotes the number of images in the database. These distance vectors are then labelled according to the relevances: those $D(Q_m, X_n)$ where X_n is relevant with respect to Q_m are labelled +1 (relevant) and the remaining ones are labelled with the class label -1 (irrelevant).

This leads to a training set, so that any classifier can be trained. This classifier should be able not only to classify distance vectors of unseen images to an image from the database, but additionally return some score or confidence with which the database images can be ranked. In the following, we present two suitable approaches to compute these scores from the distance vectors.

4.2 Logistic Regression Feature Weights for Image Retrieval

To obtain suitable feature weights, a logistic model is promising, because it is suited to combine features of different types.

As features f_i for the logistic model we choose the distances between the i -th feature of the query image Q and the database image X_n :

$$f_i(Q, X_n) := d_i(Q_i, X_{ni}).$$

To allow for modelling prior probabilities, we include a constant feature $f_{i=0}(Q, X_n) = 1$. Then, the scores $S(Q, X_n)$ from Eq. (1) are replaced by the posterior probability for class +1 and the ranking and combination of several query images is done as before:

$$\begin{aligned} S(Q, X_n) &:= p(+1|Q, X_n) \\ &= \frac{\exp[\sum_i \lambda_{+1i} f_i(Q, X_n)]}{\sum_{k \in \{+1, -1\}} \exp[\sum_i \lambda_{ki} f_i(Q, X_n)]} \\ &= \frac{1}{1 + \exp(\sum_i \lambda_i f_i(Q, X_n))} \text{ with } \lambda_i = \lambda_{-1i} - \lambda_{+1i} \end{aligned} \quad (3)$$

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 λ_{+1i} and λ_{-1i} .

We train the λ of the logistic model from Eq. (3) using the GIS algorithm.

4.3 Support Vector Machine Scoring

Since dividing given distances into “relevant” and “irrelevant” is a two-class problem, it is quite natural to employ a support vector machine (SVM) [16].

SVMs are, contrary to logistic models described above, not a probabilistic method providing class-posterior probabilities to base the classification decision upon, but directly predict the label of the observation. An SVM commonly discriminates between two classes: -1 and $+1$, using the decision rule to classify an unseen observation X :

$$X \mapsto \hat{c}(X) = \text{sgn} \left\{ \sum_{v_i \in \mathcal{S}} \alpha_i K(X, v_i) + \beta \right\} \quad (4)$$

where K is a kernel function, $\mathcal{S}$ is the set of support vectors v_i , and the α_i are the corresponding weights, β is a bias term.

Considering the distance vector D , as described above as feature vectors, it is possible to rank images using the distances to the separating hyperplane. That is, given the distance vector $D(Q_m, X_n)$, by computing the distance

$$d(D(Q_m, X_n)) = \sum_{v_i \in \mathcal{S}} \alpha_i K(D(Q_m, X_n), v_i) + \beta, \quad (5)$$

it is possible to compute a score $S(q, X) = \exp(d)$, which can then be used to replace the score of Eq. 1.

Since the number of “relevant” distance vectors given the photographic retrieval task is small compared to the number of “irrelevant” ones we randomly select a subset of “irrelevant” distance vectors to have the same number of distance vectors for both classes. Informal experiments have shown that using far more vectors from one class than from the other decreases the performance.

Table 1. Overview of our submitted results, competing submissions, and improvements presented in this work. “emp” denotes empirically determined weights, “ME” denotes maximum-entropy(logistic) scoring, and “SVM” denotes support vector scoring.

submission	with text.	MAP	comment
average-NT	no	0.07	with query expansion
RWTH-FIRE-NT-emp	no	0.08	
RWTH-FIRE-ME-NT-20000	no	0.11	
best-NT	no	0.19	with query expansion
average(monolingual English)	yes	0.14	
RWTH-FIRE-emp	yes	0.20	
RWTH-FIRE-ME-500	yes	0.20	with query expansion
best(monolingual English)	yes	0.32	
SVM-rbf-NT-withsift	yes	0.13	this work
FIRE-emp-withsift	yes	0.20	this work
SVM-linear	yes	0.21	this work
SVM-rbf	yes	0.25	this work

5 Experimental Results

In total, we submitted nine runs to the photographic retrieval task, five using textual and visual information jointly and four runs using only visual information. As can be seen in Table 1, textual information (monolingual English) greatly helps to achieve a better retrieval result, which was to be expected. In the visual-only runs, logistic regression also clearly helps to improve the results. It should be noted that 90.6% of the submissions to the photographic retrieval task used query expansion, which we did not use at all.

In the following, we show how to further improve our results.

5.1 SIFT Features

In Table 2(a) an overview of the performance of clustered SIFT histograms using different numbers of clusters is given. It can be seen that the performance increase correlates strongly with the number of clusters. However, using more than 1024 clusters did not lead to any more improvement. It can also be seen that the SIFT Features perform significantly better than global colour histograms, which usually perform quite well on this type of images. This is a strong indicator that the SIFT features capture important local information. Table 3 shows the results of adding SIFT histograms to our system, where they led to slight improvements using manually tuned weights but did not help significantly when other strong features were present, or a strong classifier was used.

5.2 SVM Scoring

The SVM Scoring approach presented in section 4.3 helped increase the MAP of our submissions by up to 25% relative. Even using linear kernels for the SVM, the performance increases compared to the logistic approach. The best

Table 2. (a) Performance using clustered SIFT histograms with different numbers of clusters. (b) Different feature combination strategies compared.

(a)		(b)		
number of clusters	map	scoring	train(2006)	test(2007)
colour histogram	0.022			
256	0.027	emp	0.1625	0.1969
512	0.039	ME	0.1479	0.1974
1024	0.046	SVM-linear	0.1581	0.2080
2048	0.042	SVM-rbf	0.2091	0.2460

Table 3. Performance increase using SIFT features in combination with other visual features and in combination with other visual features and text

	features used	unweighted	emp	ME	SVM-rbf
no text	baseline	0.0840	0.1017	0.1122	0.1282
	+sift	0.0870	0.1100	0.1110	0.1302
with text	baseline	0.1260	0.1946	0.1970	0.246
	+sift	0.1290	0.2015	0.1970	0.241

results were achieved using an RBF-kernel with parameters estimated on the queries of the 2006 photographic retrieval task, which were used as development data. An overview of the results using the different scoring approaches is given in Table 2(b) which compares the unweighted baseline to hand-tuned linear weights, logistic scoring and SVM-scoring.

6 Conclusion

In this work, we described our approach to the photographic retrieval task of the ImageCLEF2007 evaluation. It can be seen that, for the given task, textual information is crucial to obtain good retrieval accuracy. Among the visual features, the clustered SIFT histograms perform better than other widely used features if used on their own, nonetheless in combination with other features they only lead to minor improvements.

Additionally, we presented an SVM-based approach for ranking retrieval results. This method outperforms our logistic regression scoring method by up to 25%, relatively.

For the future, the impact of user interaction is an important research topic. In particular it might be interesting to investigate discriminative machine learning techniques to learn good feature weights from user interaction.

Acknowledgement

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

References

1. Deselaers, T., Weyand, T., Ney, H.: Image retrieval and annotation using maximum entropy. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 725–734. Springer, Heidelberg (2007)
2. Setia, L., Ick, J., Burkhardt, H.: Svm-based relevance feedback in image retrieval using invariant feature histograms. In: IAPR Workshop on Machine Vision Applications, Tsukuba Science City, Japan (2005)
3. Yavlinski, A., Pickering, M.J., Heesch, D., Rüger, S.: A comparative study of evidence combination strategies. In: ICASSP 2004, Montreal, Canada, vol. 3, pp. 1040–1043 (2004)
4. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Proceedings of the CLEF 2007 Workshop, Budapest, Hungary. LNCS, vol. 5152, pp. 433–444. Springer, Heidelberg (2008)
5. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr benchmark: A new evaluation resource for visual information systems. In: LREC 2006 OntoImage 2006: Language Resources for Content-Based Image Retrieval, Genoa, Italy (2006)
6. 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)
7. Puzicha, J., Rubner, Y., Tomasi, C., Buhmann, J.: Empirical evaluation of dissimilarity measures for color and texture. In: ICCV 1999, Corfu, Greece, vol. 2, pp. 1165–1173 (1999)
8. Deselaers, T.: Features for image retrieval. Master's thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany (2003)
9. Siggelkow, S.: Feature Histograms for Content-Based Image Retrieval. PhD thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany (2002)
10. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. *IEEE Transaction on Systems, Man, and Cybernetics* 8, 460–472 (1978)
11. Squire, D.M., Müller, W., Müller, H., Raki, J.: Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In: SCIA, Kangerlussuaq, Greenland, pp. 143–149 (1999)
12. Deselaers, T., Keysers, D., Ney, H.: Discriminative training for object recognition using image patches. In: CVPR 2005, San Diego, CA, vol. 2, pp. 157–162 (2005)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)
14. Dorkó, G., Schmid, C.: Object class recognition using discriminative local features. Rapport de recherche RR-5497, INRIA - Rhone-Alpes (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.) DAGM 2006. LNCS, vol. 4174, pp. 202–211. Springer, Heidelberg (2006)
16. Schölkopf, B.: Support Vector Learning. Oldenbourg Verlag, Munich (1997)

MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval

Julio Villena-Román^{1,3}, Sara Lana-Serrano^{2,3},
José Luis Martínez-Fernández^{1,3}, and José Carlos González-Cristóbal^{2,3}

¹ Universidad Carlos III de Madrid

² Universidad Politécnica de Madrid

³ DAEDALUS – Data, Decisions and Language, S.A.

jvillena@it.uc3m.es, slana@diatel.upm.es,
joseluis.martinez@uc3m.es, josecarlos.gonzalez@upm.es

Abstract. This paper describes the participation of MIRACLE research consortium at the ImageCLEF Photographic Retrieval task of ImageCLEF 2007. For this campaign, the main purpose of our experiments was to thoroughly study different merging strategies, i.e. methods of combination of textual and visual retrieval techniques. Whereas we have applied all the well known techniques which had already been used in previous campaigns, for both textual and visual components of the system, our research has primarily focused on the idea of performing all possible combinations of those techniques in order to evaluate which ones may offer the best results and analyze if the combined results may improve (in terms of MAP) the individual ones.

Keywords: Linguistic Engineering, Information Retrieval, text retrieval, image retrieval, merging strategies, boolean operators, score computation.

1 Introduction

MIRACLE is a research consortium formed by research groups of three different Spanish universities (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a private company founded as a spin-off of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in many different tracks and tasks, including the main bilingual, monolingual and cross lingual tasks as well as ImageCLEF, QA, WebCLEF and GeoCLEF tracks.

This paper describes our participation [1] at the ImageCLEF Photographic Retrieval task [2] of ImageCLEF 2007. The main purpose of our experiments was to thoroughly study the different merging strategies, i.e. methods of combination of visual and textual techniques. Our approach is based on the combination of the text-based retrieval techniques using Lucene/Xapian [3] [4] with the GIFT/FIRE [5] [6] content-based retrieval and our research has focused on the idea to perform all possible combinations of those techniques to evaluate which ones offer the best results.

2 System Description

According to our experience in previous campaigns, a special effort was made to develop a flexible system composed of a set of small components that can be easily added in different configurations, so as to be able to execute a large number of runs that exhaustively cover all the combinations of the different techniques. Our system is built up from three main components [1] that are executed sequentially to build the final result set.

The first component is the **textual** (text-based) **retrieval module**, which indexes the IAPR TC-12 [2] image descriptions to look for those descriptions that are more relevant to the text of the topic. Since MIRACLE has taken part in CLEF, different linguistic and statistical techniques have been used in the text-based part of the different tasks [7] [8]. This year, the main goal was to make an exhaustive study of these diverse methods, combining them in all possible ways and testing the results achieved. The list of components includes proper noun detection, linguistic analyzer, stemming, stopword detection, semantic expansion, indexing and retrieval. Two different search engines have been used, Lucene [3] and Xapian [4].

The second component is the **visual** (content-based) **retrieval module**, that provides the IAPR TC-12 images which are more similar to the given topic images. For this part of the system, we used two freely available CBIR systems: GIFT (GNU Image Finding Tool) [5] and FIRE (Flexible Image Retrieval Engine) [6] [9].

Finally, the most important component is **merging module**, which combines the outputs of the two previous subsystems to provide the final results. The textual and visual result lists are merged by applying different techniques [1], which are characterized by an operator and a metric for computing the relevance. Table 1 shows the defined operators: union (OR), intersection (AND) and external join (LEFT/RIGHT JOIN). Operators select which images from the two original sets are part of the final result set. Then, merged results are reranked by computing a new relevance measure value based on their corresponding values, using different metrics shown in Table 2.

Table 1. Combination operators

Operators	
OR	$A \cup B$
AND	$A \cap B$
LEFT	$(A \cup B) \cup (A - B)$
RIGHT	$(A \cup B) \cup (B - A)$

Table 2. Score computation metrics

Metrics	Score
max	$\max(a, b)$
min	$\min(a, b)$
avg	$\text{avg}(a, b)$
	$\max(a, b) +$
mm	$\min(a, b) * \frac{\min(a, b)}{\max(a, b) + \min(a, b)}$

3 Experiments and Results

Experiments are defined by the choice of different combinations of the previously described modules, operators and score computation metrics. All experiments are fully automatic, avoiding any manual intervention. None of the experiments incorporates relevance feedback, due to time constraints. We finally submitted a wide set of

Table 3. Best-ranked experiments

	Type	MAP	P10	P20	P30	RelRet ⁽¹⁾
TxtXaTIDELONPS	Textual, mono	0.1995	0.3033	0.2783	0.2528	1892
MulESXaNPS	Textual, Spanish	0.1917	0.2983	0.2633	0.2506	1896
VisG0F0ANDmin	Visual	0.1079	0.3383	0.2400	0.1956	801
Mix2LEFTmm	Mixed	0.2244	0.4450	0.3617	0.3067	1888

⁽¹⁾ out of 3,416 relevant images.

experiments: 110 multilingual textual (text-based) runs, 22 visual (content-based) runs and 21 mixed runs (using a combination of both) [1]. Results are summarized in Table 3.

In general, textual experiments show [2] that Lucene turns out to give worse results than Xapian. A possible explanation could be that Xapian is based on the probabilistic IR model, which is more suitable when dealing with short documents, as the image descriptions. However, this conclusion has to be further studied.

Moreover, no strategy for merging results from both search engines have led to any improvement in MAP with respect to the baseline experiments, neither with the OR operator (which was supposed to increase the number of results at the expense of a loss of precision) nor the AND operator (which was supposed to increase precision).

The best results correspond to Spanish, our mother tongue in which we have a strong expertise, with a MAP similar to the monolingual experiments. We are negatively surprised by the low precision with French and German languages that may be attributed to a deficient stemming module. This issue also has to be further studied.

MAP is very low for the visual-based experiments due to the complexity of the visual retrieval task in this domain. The best experiment is the combination of results with GIFT and FIRE, and it is better than using GIFT alone. However, we regretfully did not execute the experiment with FIRE alone, which would have been shown whether the combination of both systems improves the final result. Also note that, although the number of relevant images retrieved using AND is lower than when OR is used, the precision is higher.

The best experiment uses a combination of all textual and visual results obtained in the previous experiments with the LEFT operator along with the mm (max-min) scoring metric. Moreover, the mm metric clearly outperforms the others even with the same operator over the same result sets (Mix2LEFT experiments). The most interesting conclusion is that this merging strategy is successful as the MAP is higher than in both the textual and visual partial experiments with which the experiment is built.

4 Conclusions and Future Work

Our main interest was not in experiments where only text or image content is used in the retrieval process. Instead, our challenge was to test whether the text-based image retrieval could improve the analysis of the content of the image, or vice versa. The most interesting conclusion to be drawn is that merging strategies are successful as our best mixed experiment outperforms both the textual and visual experiments in

which it is based, using the LEFT operator for the combination along with the mm (max-min) metric for computing the relevance. This result shows that our initial hypothesis was right. Our combination of a “black-box” search combining publicly accessible content-based retrieval engines with a text-based search has turned out to provide better results than other presumably “more complex” techniques. This simplicity may be a good starting point for the implementation of a real system.

Acknowledgements. This work has been partially supported 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; and by the Madrid’s R&D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

1. Villena-Román, J., Lana-Serrano, S., Martínez-Fernández, J.L., González-Cristóbal, J.C.: MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval. In: Working Notes of the Cross Language Evaluation Forum 2007, Budapest, Hungary (2007)
2. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
3. Apache Lucene project, <http://lucene.apache.org>
4. Xapian: an Open Source Probabilistic Information Retrieval library, <http://www.xapian.org>
5. GIFT: GNU Image-Finding Tool, <http://www.gnu.org/software/gift/>
6. FIRE: Flexible Image Retrieval System, <http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html>
7. Martínez-Fernández, J.L., Villena-Román, J., García-Serrano, A.M., Martínez-Fernández, P.: MIRACLE team report for ImageCLEF IR in CLEF 2006. In: Proceedings of the Cross Language Evaluation Forum 2006, Alicante, Spain (2006)
8. Martínez-Fernández, J.L., Villena-Román, J., García-Serrano, A.M., González-Cristóbal, J.C.: Combining Textual and Visual Features for 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. Springer, Heidelberg (2006)
9. Deselaers, T., Keysers, D., Ney, H.: FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In: Peters, C., et al. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 688–698. Springer, Heidelberg (2004)

Using an Image-Text Parallel Corpus and the Web for Query Expansion in 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. ImageCLEF2007 photo task is different from those of the previous years in two aspects. The caption field in the image annotations and the narrative field in the text queries are removed, and the example images in the visual queries are also removed from the image collection. In the new definition, the information that can be employed is less than before. Thus matching query words and annotations directly is not feasible. This paper explores the web to expand queries and documents. The experiments show that query expansion improves the performance 16.11%, however, document expansion brings in too much noise and the performance decreases 28.24%. The media mapping method based on an image-text parallel corpus is regarded as query expansion. The results of the formal runs show this method performs the best. Compared with the performance of the models without expansion, the MAP improves about 86.69%~143.12%. Integration of the external and the internal resources gains no benefits in the further experiments.

1 Introduction

Image retrieval becomes more important since large scale image data are available on the web. In ImageCLEFphoto task, each topic, which is composed of a text query and a visual query, simulates the information need of users. In the previous years [1] [2], a text query includes topic and narrative fields in several different languages and a visual query includes two or three example images in the image data set. Each image in the image collection is annotated with title, location, date, notes and a detail caption. The task definitions of ImageCLEFphoto2007 [3] are changed in two aspects. First, the caption field in the image annotations and the narrative field in the queries are removed. These changes aim to reflect the information needs of the real world, i.e., image annotations and queries are usually short and rough. Second, the example images of the visual queries do not belong to the image collection. This change reflects that users may use their own photos as examples rather than images in the collection.

When the caption field in the image annotation is removed, matching query words and image annotations becomes more challenging than before. We will explore an external resource, e.g., the web, to expand queries and documents. Through a web

* Corresponding author.

search engine such as Google, we can retrieve relevant web pages, and use them for expansion. Compared with query expansion of using pseudo relevance feedback in the corpus, the outside resource may bring in information that the target corpus may not have. However, the information retrieved from the web may contain noise at the same time. How to filter out noise is an important issue. In this paper, we restrict the search space to some kinds of web sites, e.g., Wikipedia, and investigate if it is helpful for retrieval.

In addition to external resources, we employ internal resources such as an image-text parallel corpus, i.e., the target collection itself. Under such a trans-media corpus, two approaches – say, media mapping [4] and a trans-media dictionary [5] were proposed before. Media mapping approach, which can be regarded as a kind of pseudo relevance feedback across different media, has better performance than trans-media dictionary approach. We will employ media mapping to ImageCLEFphoto 2007, examine its performance in the new definitions, and analyze if the integration of external and internal resources is helpful.

This paper is organized as follows. Section 2 introduces the three methods we explore. Section 3 specifies official and unofficial runs we design. Section 4 shows and discusses the experiment results. Finally, we conclude the remarks in Section 5.

2 Methods

Three methods including query (document) expansion using the web and query expansion with media mapping via a cross-media corpus are presented in the following subsections.

2.1 Query Expansion Using the Web

Queries and image annotations are both short in ImageCLEFphoto 2007. In this situation, we plan to expand the given queries and get more information. Several previous experiments have shown that query expansion using pseudo relevance feedback is very useful in this task. In this paper, we employ outside resource like the web for query expansion and analyze if it can bring in useful information.

The best way to access the web is through a web search engine like Google. We submit a text query to retrieve relevant web pages. Because the language of a text query may be different from the language of image annotations, we have to introduce the language translation mechanism. There are two alternative ways to deal with this problem. The first is to submit a text query to the web search engine directly and then to translate the retrieved web pages into target language. The second is to translate a text query into target language and then to submit the translated query to the web search engine. The drawback of the first approach (i.e., translation after retrieval) is the cost to translate all the web pages we get. In contrast, the translation cost of the second approach (i.e., translation before retrieval) is relatively low. However, when there are named entities in queries, the second approach may get wrong translation and thus the retrieved web pages may be unrelated to the original query.

In the experiments, we adopt the approach of translation before retrieval. Next, we have to select words from the retrieved results to expand the given query. The selection mechanism can filter out noisy information, but it may also lose some useful

information. Here, we adopt the simplest way, i.e., to employ the top ranked snippets to expand our query. For the issue of noise, we limit the websites we access to encyclopedia-based ones, e.g., Wikipedia, by adding a web site name as an extra query term when submitting a query to a web search engine.

2.2 Document Expansion Using the Web

Direct keyword matching may not be workable after query expansion if the relevant documents do not mention the words in the expanded query. There are two alternative ways to deal with this problem. First, we can expand a query with the words appearing in a document. Query expansion using relevance feedback belongs to this type. Second, we may expand the documents.

In this paper, we explore the document expansion using the web. This method is similar to the one used in query expansion. We consider the title field of an image annotation as a query, and submit it to the web search engine to get the top ranked snippets to expand documents. Because documents are in target language, language translation is not necessary during document expansion. That is the major difference from query expansion.

Although document expansion avoids translation errors, expanding too many words may introduce noise. In document expansion, we restrict the selection scope as follows. Only those words nearby the words in the title field of image annotations are considered as candidates. We set a window size (e.g., 5) in the experiments.

Besides the above noise issue, document expansion has a logical problem. Assume the word “animal” is in title field of an image annotation. When we expand hyponym words such as “tiger”, “cat”, “dog”, etc., we do not know which animals are actually mentioned in that image. If the image talks about “rabbit”, the wrong expansion may introduce erroneous terms.

2.3 Media Mapping with an Image-Text Parallel Corpus

Media mapping method [4][6] regards the target collection as an image-text parallel corpus, and employs such an intermedia to translate a visual query into a text one, and vice versa. The intermedia link two kinds of media (i.e., text and image) in this paper. Media mapping method can be seen as relevance feedback across different media and used in query expansion. In ImageCLEFphoto2006, we created a new query using media mapping and merged the results of the new query and the original visual query.

In ImageCLEFphoto 2007, we regard the media mapping as query expansion in the following way. First, we submit a visual query to a content-based information retrieval (CBIR) system. Because images and the corresponding text annotations are parallelized in the collection, we then rerank the top n returned images by using a text query. Finally, the text annotations of the top m images are added to the original text query. We submit the expanded query to a text retrieval system and get the final result.

We can compare the results of query expansion using the web and the media mapping with intermedia. In addition, we can examine the feasibility of the media mapping method in the new task definitions. There are two new challenges. First, retrieving the related images in the intermedia via visual query becomes harder. In the past, the visual queries are images in the image collection, so that they always appear

in the top of the returned images. Second, the caption field in an image annotation has been removed beforehand, so that the text information we can get from the image counterparts is less than the one in the tasks of previous years. We are interested in if the media mapping method is still workable in the new environment.

3 Experiments

We submitted 27 official runs including 18 cross-lingual runs for eight different languages, 8 mono-lingual runs for three different languages, and 1 run for visual query only. All the queries with different source languages were translated into target language (e.g., English) using SYSTRAN system. We adopted Okapi with BM25 formula for text retrieval.

The experiments consider the following issues. First, we want to check if the retrieved web pages can bring in new information for query expansion. We examine the expanded words when the recall is improved. Second, we compare the results of query expansion runs that limit or do not limit the web sites. Third, we want to check the effects that document expansion achieves. The runs using both query expansion and document expansion are also checked. Fourth, we examine the performance of media mapping method. Then, we employ both media mapping and the web for query expansion, and check if the web can bring in new information that media mapping cannot do. Some of the above issues are verified in the official runs, while some are done in the unofficial runs.

Our official runs are described as follows. A run is named by the format *SourceLanguage-TargetLanguage-Automatic-FeedBack-Media*, where DE (German), ES (Spanish), EN (English), FR (French), JP (Japanese), RU (Russian), ZHT (Traditional Chinese), ZHS (Simplified Chinese), AUTO (Automatic), NOFB (No Feedback), TE (Document Expansion), FBQE (Feedback and Query Expansion), TXT (Text), IMG (Image), and TXTIMG (Text and Image).

1. 8 cross-lingual runs that use text query only, and do not consider query expansion:
ES-EN-AUTO-NOFB-TXT, FR-EN-AUTO-NOFB-TXT, RU-EN-AUTO-NOFB-TXT,
PT-EN-AUTO-NOFB-TXT, JA-EN-AUTO-NOFB-TXT, IT-EN-AUTO-NOFB-TXT,
ZHT-EN-AUTO-NOFB-TXT, ZHS-EN-AUTO-NOFB-TXT
2. 3 mono-lingual runs that use text query only, and do not consider query expansion:
EN-EN-AUTO-NOFB-TXT, ES-ES-AUTO-NOFB-TXT, DE-DE-AUTO-NOFB-TXT
3. 3 mono-lingual runs that adopt the media mapping method for query expansion:
ES-ES-AUTO-FBQE-TXTIMG, EN-EN-AUTO-FBQE-TXTIMG,
DE-DE-AUTO-FBQE-TXTIMG
4. 8 cross-lingual runs that use the media mapping method for query expansion:
PT-EN-AUTO-FBQE-TXTIMG, ES-EN-AUTO-FBQE-TXTIMG,
RU-EN-AUTO-FBQE-TXTIMG, IT-EN-AUTO-FBQE-TXTIMG,
ZHT-EN-AUTO-FBQE-TXTIMG, ZHS-EN-AUTO-FBQE-TXTIMG,
JA-EN-AUTO-FBQE-TXTIMG, FR-EN-AUTO-FBQE-TXTIMG
5. 2 runs that expand query with the web, but do not consider document expansion:
EN-EN-AUTO-QE-TXT-TOPIC, ZHT-EN-AUTO-QE-TXT-TOPIC
6. 2 runs that expand document with the web, but do not consider query expansion:
EN-EN-AUTO-TE-TXT-CAPTION, ZHT-EN-AUTO-TE-TXT-CAPTION

- 7. 1 run that use visual query and the media mapping method only:
IMG-EN-AUTO-FB-TXTIMG
Some unofficial runs are described as follows.
- 1. 2 runs that consider both query expansion and document expansion:
EN-EN-AUTO-TE-QE-TXT, ZHT-EN-AUTO-TE-QE-TXT
- 2. 2 runs that expand query with the web and limit the search space:
EN-EN-AUTO-QE-WIKI-TXT, ZHT-EN-AUTO-QE-WIKI-TXT
- 3. 2 runs that use both the web and the media mapping for query expansion:
EN-EN-AUTO-QE-FBQE-TXTIMG, ZHT-EN-AUTO-QE-FBQE-TXTIMG.

4 Results and Discussions

In the first set of experiments, we use the top one snippet returned by Google to expand the text queries. Table 1 shows the results of runs EN-EN-AUTO-QE-TXT-TOPIC, ZHT-EN-AUTO-QE-TXT-TOPIC, EN-EN-AUTO-NOFB-TXT, and ZHT-EN-AUTO-NOFB-TXT. In both cross-lingual and mono-lingual cases, the performance of systems with query expansion is better than that without query expansion. After expansion, both recall and precision are improved. In the original expectation, precision is decreased since we do not apply any strategies to filter noise.

Table 1. Results of models with/without query expansion

Query Language- Document Language	Evaluation Metric	Query Expansion Using the Web	Query Without Expansion
Traditional Chinese- English	MAP	0.1225 (+16.11 %)	0.1055
	Recall	0.4461 (+18.14 %)	0.3776
English-English	MAP	0.1577 (+7.57 %)	0.1466
	Recall	0.5439 (+14.84%)	0.4736

In the second set of experiments, we compare the results of query expansion with and without limiting the search space. Table 2 shows that the performance does not change very much after restricting the web sites for cross-lingual retrieval. The performance even has a little decrease in mono-lingual runs. We find that restrictive access may retrieve unrelated pages in some cases.

The third set of experiments aims to evaluate the effects of document expansion. Table 3 summarizes the results. Document expansion does not take positive effects. In both cross-lingual and mono-lingual runs, recall and MAP are decreased when document expansion is introduced no matter whether query expansion is employed or not. The major reason may be that the strategy brings in too much noise.

The fourth set of experiments examines the performance of the media mapping method in the new definitions. The results are shown in Table 4. Total 8 cross-lingual runs and 3 mono-lingual runs are tested. Media mapping achieves very good performance. Compared with the performance of the models without expansion, the MAP improves about 86.69%~143.12%. In last year, media mapping improves the performance about 71%~119%. This result shows that media mapping method is robust under different task definitions.

Table 2. Results of models with and without limiting the search space

Run Name (cross-lingual/mono-lingual)	Limitation	Recall	MAP
ZHT-EN-AUTO-QE-TXT-TOPIC (cross-lingual)	No	0.4461	0.1225
ZHT-EN-AUTO-QE-WIKI-TXT (cross-lingual)	Yes	0.4713	0.1290
EN-EN-AUTO-QE-TXT-TOPIC (mono-lingual)	No	0.5439	0.1577
EN-EN-AUTO-QE-WIKI-TXT English (mono-lingual)	Yes	0.5102	0.1330

Table 3. Results of models using or not using document expansion

Runs Name (cross-lingual/mono-lingual)	Document Expansion	Query Expansion	Recall	MAP
ZHT-EN-AUTO-NOFB-TXT (cross)	No	No	0.3776	0.1055
ZHT-EN-AUTO-TE-TXT-CAPTION (cross)	Yes	No	0.3050	0.0757
ZHT-EN-AUTO-QE-TXT-TOPIC (cross)	No	Yes	0.4461	0.1225
ZHT-EN-AUTO-TE-QE-TXT (cross)	Yes	Yes	0.3562	0.0799
EN-EN-AUTO-NOFB-TXT (mono)	No	No	0.4736	0.1466
EN-EN-AUTO-TE-TXT-CAPTION (mono)	Yes	No	0.3729	0.1154
EN-EN-AUTO-QE-TXT-TOPIC (mono)	No	Yes	0.5439	0.1577
EN-EN-AUTO-TE-QE-TXT (mono)	Yes	Yes	0.4156	0.1203

Table 4. Results of using the media mapping as query expansion

Query Language-Document Language	Metric	Query Expansion using Media Mapping	Without Expansion
Traditional Chinese-English	MAP	0.2565 (+143.12 %)	0.1055
	Recall	0.6405 (+69.62 %)	0.3776
Simplified Chinese-English	MAP	0.2565 (+143.12 %)	0.1055
	Recall	0.6405 (+69.62 %)	0.3776
Portuguese-English	MAP	0.2820 (+109.35 %)	0.1347
	Recall	0.6733 (+51.81 %)	0.4435
Spanish-English	MAP	0.2785 (+96.12 %)	0.1420
	Recall	0.6718 (+50.89 %)	0.4452
Russian-English	MAP	0.2731 (+100.66 %)	0.1361
	Recall	0.6738 (+46.54 %)	0.4598
Italian-English	MAP	0.2705 (+130.60 %)	0.1173
	Recall	0.6481 (+67.59 %)	0.3867
French-English	MAP	0.2669 (+95.96 %)	0.1362
	Recall	0.6651 (+61.74 %)	0.4112
Japanese-English	MAP	0.2551 (+117.29 %)	0.1174
	Recall	0.6507 (+57.55 %)	0.4130
English-English	MAP	0.2737 (+86.69 %)	0.1466
	Recall	0.6812 (+43.83 %)	0.4736
Spanish-Spanish	MAP	0.2792 (+92.02 %)	0.1454
	Recall	0.6282 (+32.30 %)	0.4748
German-German	MAP	0.2449 (+128.87 %)	0.1070
	Recall	0.5790 (+82.64 %)	0.3170

Tables 1 and 4 conclude that media mapping with an image-text parallel corpus and query expansion using the web are very useful, and the former is better than the latter. The last set of experiments checks if integrating the internal and the external resources gets better performance. Table 5 shows that such an integration does not have positive effects. MAP is decreased when the web is used. It may be due to that the external resource (i.e., the web) has more noise than the internal resource (i.e., the image-text parallel corpus).

Table 5. Results of the models using both media mapping and the web

Run	Media Mapping	the Web	Recall	MAP
ZHT-EN-AUTO-FBQE-TXTIMG	Yes	No	0.6405	0.2565
ZHT-EN-AUTO-QE-FBQE-TXTIMG	Yes	Yes	0.6533	0.2255
EN-EN-AUTO-FBQE-TXTIMG	Yes	No	0.6812	0.2737
EN-EN-AUTO-QE-FBQE-TXTIMG	Yes	Yes	0.6738	0.2442

In the above sets of experiments, we compare the performance of different kinds of approaches. Media mapping with an image-text parallel corpus is the best of all. Table 6 summarizes the ranks of our official runs with media mapping approach in ImageCLEFphoto2007. Each row lists the language pair, total submitted runs and the rank of our runs. Compared with the runs of different participants, media mapping approach performs quite well in different language pairs. Except English mono-lingual and Simplified Chinese-English cross-lingual runs, our system ranks number 1 in the rest of official runs we submitted. That shows the robustness of our media mapping approach in integrating text and visual information.

Table 6. Ranks of official runs using media mapping approach in ImageCLEFphoto2007

Mono-Lingual Run/Cross-Lingual Run	Total Submitted Runs	Rank
English → English	142	8
German → German	30	1
Spanish → Spanish	15	1
Simplified Chinese → English	23	2
Tradition Chinese → English	1	1
French → English	21	1
Italian → English	10	1
Japanese → English	6	1
Portuguese → English	9	1
Russian → English	6	1
Spanish → English	9	1

5 Conclusion

This paper explores the use of the web for query and document expansion. The experiments show that the named entities expanded from the web are useful. Limiting the search web sites to Wikipedia seems not to improve the performance and may filter out some related webs. Document expansion brings in too much noise, so that

the performance decreases 28.24%. Regarding media mapping as query expansion improves the retrieval performance very much. It shows the robustness of media mapping method even the new task definition is more challenging than before. Integrating both the web and an image-text parallel corpus for query expansion cannot improve the performance.

Acknowledgments. Research of this paper was partially supported by National Science Council, Taiwan, and Excellent Research Projects of National Taiwan University, under the contracts 95-2221-E-002-334 and 96R0062-AE00-02.

References

1. Clough, P., Sanderson, M., Müller, H.: The CLEF 2004 Cross-Language Image Retrieval Track. In: Peters, C., Clough, P., 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., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersch, W.: The CLEF 2005 Cross-Language Image Retrieval Track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
3. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEFphoto 2007 Photographic Retrieval Task. In: Nardi, A., Peters, C. (eds.) Working Notes of the 2007 CLEF Workshop (2007)
4. Chang, Y.C., Chen, H.H.: Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 625–632. Springer, Heidelberg (2007)
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* 43, 488–502 (2007)
6. Chen, H.H., Chang, Y.C.: Language Translation and Media Transformation in Cross-Language Image Retrieval. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 350–359. Springer, Heidelberg (2006)

SINAI System: Combining IR Systems at ImageCLEFPhoto 2007

M.Á. García-Cumbreras, M.C. Díaz-Galiano, M.T. Martín-Valdivia,
A. Montejo-Ráez, and L.A. Ureña-López

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

Abstract. This paper describes the SINAI team participation in the ImageCLEFPhoto 2007 campaign. This year we have developed a system that combines the document lists retrieved by two Information Retrieval systems (Lemur and JIRS). Online machine translators have been used for the bilingual experiments. The results obtained show that if we only use title text our system works bad. Because of the low MAP values fusion method does not improve the results.

1 Introduction

This is the third participation of the SINAI research group at the ImageCLEF campaign [1]. We have participated in the AdHoc task [2] and in the medical task.

The AdHoc task involves retrieving relevant images using the text associated to each image query. As a cross-language retrieval task, multilingual image retrieval based on query translation can achieve higher performance than monolingual retrieval.

This year, a new Information Retrieval (IR) module has been tested. This module works with two different IR systems and the final relevant list is the result of the combination of both IR lists (voting system). The Machine Translation Module developed last year has been updated and used for the bilingual task. English, Spanish, French, Italian and Portuguese are the languages used this year.

Given a multilingual query, the goal of the Image CLEF Photographic task is to find as many relevant images as possible from an image collection.

The proposal is to compare results with and without pseudo-relevant feedback (PRF), with or without query expansion, using different methods of query translation or using different retrieval models and weighting functions.

The following sections describe the SINAI system, and our experiments are detailed. Finally, conclusions and further work are presented.

2 System Description

2.1 Collection Preprocessing

The dataset used is the IAPR collection. The IAPR TC-12 image collection consists of 20,000 images taken from different locations around the world and

comprises a varying cross-section of still natural images. It includes pictures of a range of sports and actions, photographs of people, animals, cities, landscapes and many others of contemporary life.

The collections have been preprocessed using stopwords removal and the Porter's stemmer.

The dataset has been indexed using both IR systems, namely, Lemur¹ (used past years) and JIRS [3]. Java Information Retrieval System (JIRS) is a Passage Retrieval system oriented to Question Answering tasks although it can be applied as IR system.

One parameter for each experiment is the weighting function, such as Okapi or $TF \cdot IDF$. Another is the use or not of pseudo-relevance feedback (PRF).

2.2 Queries Processing

Given a set of multilingual queries, in the bilingual subtasks, the first step is its translation into English.

As translation module we have used SINTRAM (SINai TRAnslation Module), our Meta Machine Translation system that uses some online Machine Translators for each language pair, and implements some heuristics to combine the different translations [4].

We have made previous experiments using the same translation module. The best result for each language is obtained by the following translators:

- Systran for French, Italian and Portuguese queries
- Prompt for Spanish queries

Then, the original and translated English queries have been preprocessed, as usual (stopper and stemmer), and run against the IR index.

2.3 Experiments Description

In our experiments we have used English queries (monolingual) and the four following bilingual: French, Italian, Portuguese and Spanish.

Our system combines lists of relevant documents returned by Lemur and JIRS IR systems.

A simple fusion method has been implemented to obtain a simple list of relevant documents. In a first step, both lists are normalized between 0 and 1. Then, some heuristics are applied:

- *Weighting each list.* Some experiments are based on a weighting function that gives a percentage of relevance to the Lemur list another and to the JIRS list. The final score of each relevant document is calculated by the sum of each score multiplied by its weight. Finally, the documents are sorted by their final fusion score.

¹ <http://www.lemurproject.org>

- *Using a threshold.* Another heuristic filters relevant documents by a threshold value. If the score of a document is worse than this parameter then it is not included in the final list. This final list is ranked again by the score of the documents.

Figure 1 describes the architecture of our system. Each query is translated and run against the Lemur and the JIRS Information Retrieval systems. Then, several fusion methods are applied to combine both relevant documents lists.

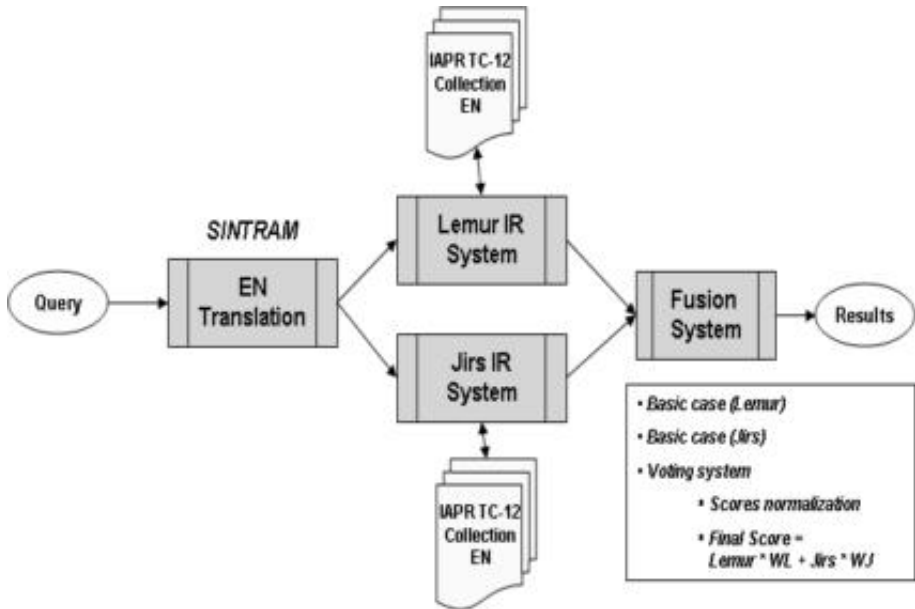


Fig. 1. Architecture of the SINAI system applied to ImageCLEFPhoto 2007

Using the AdHoc framework of ImageCLEFPhoto 2006 [5], all the described heuristics have been evaluated, in order to obtain the best configuration of parameters.

1. The Lemur baseline uses English queries and Lemur as IR system. Several weighting functions and the use or not of pseudo-relevance feedback (PRF) have been tested. The best result was obtained using Okapi as weighting function with PRF. It obtains a MAP value of 0.1672
2. The JIRS baseline uses English queries, JIRS as IR system and Okapi with PRF as weighting function. It obtains a MAP value of 0.1513
3. In the other experiments, the score of the Lemur subsystem and the JIRS one are weighted, between 0.0 and 1.0. For instance, the experiment that weights both lists in the same percentage applies the formula: $0.5 \cdot W_{lemur} + 0.5 \cdot W_{jirs}$. The best result was 0.1678 (MAP), using a weight of 0.6 for Lemur and 0.4 for JIRS.

4. To apply the second heuristic, different values, from 0.1 to 0.9, are tested as threshold. The best result was 0.1524 (MAP), obtained with a threshold=0.1.

3 Results and Discussion

We have accomplished 15 experiments: five experiments using Lemur, five using JIRS and five with the fusion of both lists.

The results obtained with each IR system (using only text, Okapi as weighting method and without expansion) and the best MAP achieved by CLEF participants for each language is shown in Table 1.

Table 1. Summary of results for the photo task: Monolingual and bilingual runs with Lemur and JIRS IR systems

Language	Experiment	IR	MAP	Best MAP
English	EN-EN-Exp2	Lemur	0.1591	0.2075
English	EN-EN-Exp1	JIRS	0.1473	0.2075
Spanish	ES-EN-Exp9	JIRS	0.1555	0.1558
Spanish	ES-EN-Exp10	Lemur	0.1498	0.1558
Portuguese	PO-EN-Exp8	Lemur	0.1490	0.1490
Portuguese	PO-EN-Exp7	JIRS	0.1350	0.1490
French	FR-EN-Exp4	Lemur	0.1264	0.1362
French	FR-EN-Exp3	JIRS	0.1195	0.1362
Italian	IT-EN-Exp5	JIRS	0.1231	0.1341
Italian	IT-EN-Exp6	Lemur	0.1198	0.1341

The results obtained with both IR systems, compared with other participants with the same configuration, are good. Only the English runs have obtained a loss of MAP of around 25%. Our best Spanish result is similar to the best one obtained. For Portuguese we have obtained the best one, and for French and Italian our results are a bit worse: only a loss of MAP of around 8%.

From these results we can conclude that the Lemur IR system works better than JIRS, but the difference is not very significant.

The results in terms of MAP are low. The experiments accomplished last year, with the same collection and same queries, gave us better results. The Table 2 show, for each language, the best result obtained in 2006, the best one obtained in 2007 and the loss of MAP obtained in 2007 (in percentage).

In 2006 title and narrative were used. In 2007 only title. All results are obtained applying query expansion and the Okapi weighting function.

After a complete analysis, the first conclusion is that if we only use the title of the query (very few words) instead of title and description, MAP results are decreased notably.

English monolingual queries obtain a loss of precision of 28%. Italian and Portuguese queries obtained better results with the new model. Spanish and French queries obtained a loss of precision around 15%.

Table 2. Comparison of results for the monolingual and bilingual runs obtained in 2006 and 2007. Last column shows the loss of MAP.

Language	MAP-2006	MAP-2007	Loss of MAP(%)
Monolingual En	0.2234	0.1591	28%
Bilingual FrEn(French)	0.1617	0.1362	15.76%
Bilingual ItEn(Italian)	0.1216	0.1341	+10.27%
Bilingual PtEn(Portuguese)	0.0728	0.1490	+104.67%
Bilingual EsEn(Spanish)	0.1849	0.1558	15.73%

The new model has worked well with the bilingual runs, but the monolingual one has decreased its results. The main reason is that the new version used of the Lemur IR system works bad than the previous one, using the same configuration.

Finally, the results obtained by applying the fusion method and the best MAP for each language is shown in Table 3.

Table 3. Summary of results for the photo task: Monolingual and bilingual runs with lists fusion

Language	Experiment	IR	MAP	Best MAP
English	EN-EN-Exp11	Fusion	0.0786	0.2075
Spanish	ES-EN-Exp15	Fusion	0.0559	0.1558
Portuguese	PO-EN-Exp14	Fusion	0.0423	0.1490
French	FR-EN-Exp12	Fusion	0.0323	0.1362
Italian	IT-EN-Exp13	Fusion	0.0492	0.1341

Fusion results have not improved the single ones. Lower MAP values decreased when we combine relevant lists. Other techniques must be used when the queries have few words.

4 Conclusions and Further Work

We have presented a system that combines document lists retrieved by two IR systems (Lemur and JIRS), and uses online translators for the bilingual experiments.

The results obtained have a low MAP, because only the title was used. Because of the low MAP values fusion method obtained poor results.

As future work, it could be interesting to develop a new robust fusion module in order to improve MAP values, and to apply a query expansion module based on Google[6], tasks on which we are already working.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id.514 granted by the University of Jaén.

References

- [1] Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2007 Ad Hoc Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2007)
- [2] Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 Photographic Retrieval Task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
- [3] 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, pp. 443–450. Springer, Heidelberg (2005)
- [4] García-Cumbreras, M.A., Ureña-López, L.A., Martínez-Santiago, F., Perea-Ortega, J.M.: BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. In: Working Notes of the 2006 CLEF Workshop (2006)
- [5] Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks. In: Working Notes of the 2006 CLEF Workshop (2006)
- [6] Martínez-Santiago, F., Montejo-Ráez, A., García-Cumbreras, M.A.: SINAI at CLEF Ad-Hoc Robust Track 2007: applying Google search engine for robust cross-lingual retrieval. In: Working Notes of the 2007 CLEF Workshop (2007)

Multimodal Retrieval by Text–Segment Biclustering^{*}

András Benczúr, István Bíró, Mátyás Brendel, Károly Csalogány, Bálint Daróczy,
and Dávid Siklósi

Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{benczur, ibiro, mbrendel, cskaresz, daroczyb, sdavid}@ilab.sztaki.hu
<http://datamining.sztaki.hu/>

Abstract. We describe our approach to the ImageCLEFphoto 2007 task. The novelty of our method consists of biclustering image segments and annotation words. Given the query words, it is possible to select the image segment clusters that have strongest cooccurrence with the corresponding word clusters. These image segment clusters act as the selected segments relevant to a query. We rank text hits by our own tf.idf-based information retrieval system and image similarities by using a 20-dimensional vector describing the visual content of an image segment. Relevant image segments were selected by the biclustering procedure. Images were segmented by graph-based segmentation. We used neither query expansion nor relevance feedback; queries were generated automatically from the title and the description words. The later were weighted by 0.1.

1 Introduction

In this paper we describe our approach to the ImageCLEF Photo 2007 evaluation campaign [1]. The key feature of our solution is to combine text based image retrieval and content based image retrieval introducing biclustering algorithm of image segments and annotation words to form interrelated clusters. Our CBIR method is based on the segmentation of the image and on the comparison of features of the segments. The biclustering algorithm is used to filter out irrelevant segments. The text retrieval system is described in [2] with two differences in using:

- A wider set of stop words including “photo”, “image” etc.;
- Heavy weight to hits in the location field but using e.g. South as location stop word.

Query terms from the topic title have ten times higher weight than the narrative terms, however, sentences containing phrase “not relevant” were automatically removed.

As our main result, we demonstrate that biclustering of image segments and annotation words additively improves retrieval performance by over 2%. In future work query expansion and feedback will be used to test whether the method can improve performance over the state of the art.

^{*} This work was supported by a Yahoo! Faculty Research Grant and by grants *MOLINGV* NKFP-2/0024/2005, NKFP-2004 project Language Miner.

2 The Content-Based Information Retrieval System

Our CBIR system relies on so called blobs, regions or segments similar to for example those of [3,4,5,6]. For each topic three sample images were given; we used the minimum of their distances from the target image for ranking. Distances were computed based on the segments of the target and sample image; for segmentation we used the code of the Felzenszwalb and Huttenlocher [7] graph-based method.

First we describe how we measure the distance between segments. For each segment we use a minimalistic 20-dimensional real valued feature vector and Euclidean distance after normalization. Out of the 20 values, 15 consist of histograms with 5 bins in each of the RGB channel and an additional 3 values contain the average intensity. The single shape information consists of the ratio of the logarithm of segment width and height. Finally the last value is the logarithm of the size in pixels.

Given the distance $\text{dist}(S, S')$ of two segments, the distance of image X to sample image I is computed from pairwise distances between pairs of segments $S(X)$ and $S(I)$ of images X and I , respectively. Since segments of I are considered as the description of the search goal, we averaged over $S_i \in S(I)$ such that for each S_i we took the closest segment from $S(X)$ as

$$\text{dist}(X, I) = \frac{1}{|S(I)|} \sum_j \min_i \{ \text{dist}(S_i, S_j) : S_i \in S(X), S_j \in S(I) \}. \quad (1)$$

3 Image Segment – Annotation Word Biclustering

Our method is special in using the annotation text to guide the CBIR via biclustering, a technique used in a wide variety of applications [8]. Our assumption is that biclusters indicate connection between the features and the text such as blue color and “pool”, white color and “snow”, black and white histogram and “black and white”. This can be used to select relevant segments of the sample image. Hence we compute an interrelated segment and word clustering together with a weight for each pair of a segment and a word cluster.

The output of segment–word biclustering is used to refine the CBIR method. In equation (1) we use only those segments of the three sample images where there is a topic title word in a text cluster with large weight for the given segment cluster. In the rare case when none of the segments is selected, we keep all as a fallback mechanism.

We used the biclustering algorithm of [8]. Let X and Y be discrete random variables that take values in the sets $\{\text{segments}\}$ and $\{\text{annotation words}\}$ respectively. Let $p(X, Y)$ denote the joint probability distribution of X and Y . Let the k clusters of X be $\{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_k\}$, and let the ℓ clusters of Y be $\{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_\ell\}$. We are interested in finding maps C_X and C_Y ,

$$C_X : \{x_1, x_2, \dots, x_k\} \mapsto \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_k\}, \quad C_Y : \{y_1, y_2, \dots, y_\ell\} \mapsto \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_\ell\}. \quad (2)$$

For brevity we write $\hat{X} = C_X(X)$ and $\hat{Y} = C_Y(Y)$ where $\hat{X}$ and $\hat{Y}$ are random variables that are a deterministic function of X and Y , respectively. Finally let $D(p \parallel q)$ denote the *Kullback–Leibler* divergence of probability distributions p and q .

Table 1. Comparison of performance of various methods evaluated by different measures

	MAP	P10	P20	P30	BPREF	GMAP	manual MAP
text + visual + bicluster	0.2238	0.3283	0.2875	0.2556	0.2003	0.0449	0.2545
text + visual	0.2076	0.3183	0.2683	0.2372	0.1924	0.0419	0.2441
text only	0.2020	0.3033	0.2492	0.2200	0.1747	0.0463	0.2295
visual + bicluster only	0.0138	0.0467	0.0433	0.0367	0.0240	0.0019	
visual only	0.0129	0.0683	0.0400	0.0317	0.0427	0.0021	

Table 2. Performance of best method as function of the CBIR weight in ranking

weight of image	1	10	100	1000	2000	5000	10000
MAP	0.2146	0.2152	0.2151	0.2238	0.2120	0.2027	0.1951

The algorithm of [8] iterates between computing segment (row) and word (column) clusters. As in iteration t of [8], the new cluster index of word y becomes

$$C_Y^{(t+1)}(y) = \operatorname{argmin}_{\hat{y}} D \left(\frac{p(X, y)}{p(y)} \middle| \middle| p(X) \cdot \frac{p(\hat{x}, \hat{y})}{p(\hat{x}) \cdot p(\hat{y})} \right), \quad (3)$$

resolving ties arbitrarily. We slightly modify this procedure for computing the new cluster index of segment x by using the 20-dimensional segment feature vector $f_1(x), \dots, f_s(x)$. We combine the Kullback-Leibler distance over the word incidence matrix with Euclidean distance in 20 dimensions as

$$C_X^{(t+1)}(x) = \operatorname{argmin}_{\hat{x}} \left\{ D \left(\frac{p(x, Y)}{p(x)} \middle| \middle| p(Y) \cdot \frac{p(\hat{x}, \hat{y})}{p(\hat{x}) \cdot p(\hat{y})} \right) + \sqrt{\sum_{i=1}^s (f_i(x) - f_i(\hat{x}))^2} \right\} \quad (4)$$

where $f_i(\hat{x})$ is the cluster average. We resolve ties arbitrarily.

4 Results

Table 1 shows the results of the text based, content based and the mixed method in our original submission. Visual only results are very poor; however when combined with text, our CBIR yields significant improvements in all measures. Surprisingly, our CBIR improves more over text only retrieval than its performance when used alone. This fact is further justified by using manually constructed text queries for the worst performing 6 topics (last column of Table 1). Table 2 shows the performance in function of the weight of the image based method when combining it with the text based query.

Table 3 shows a detailed analysis of the text only method for the topics. Best performances (57: “radio telescope”, 21: “accommodation, host family”, 10: “destination,

Table 3. Performance of text only method, some selected topics

topic	57	21	10...	...15...	...11...	...41	30	18
MAP	0.9650	0.9306	0.7852	0.4563	0.3119	0.00	0.00	0.00

Table 4. Some of the topics with the best improvement and worst deterioration when adding image similarities (top), when, in addition, using segment selection by biclustering (middle) and the overall visual improvement (bottom)

topic	11	27	15	6	17	53	32	43	48	10	8
MAP improv. by image	0.26	0.13	0.05	0.06	-0.08	-0.12	-0.16	0.08	0.12	-0.14	0.08
additional improv. by bic	0.31	0.14	0.14	0.06	0.17	0.18	0.16	-0.07	-0.16	0.06	-0.19
total MAP improvement	0.57	0.27	0.19	0.12	0.11	0.06	0.0	0.01	-0.04	-0.09	-0.11

Venezuela”) are achieved when title contains specific words that match the annotation style. Worst performance corresponds to the need for either expanding terms (41: “South America”, 18: “outside Australia” – the latter easily solved manually by negation) or understanding the visual semantics of the topic (30: “more than two beds”).

Image content with biclustering increases performance by more than 2% in average, including some topics with large improvement and only a slight deterioration for others as seen in table 4. Biclustering only sporadically deteriorates the CBIR performance. We see the largest improvement from topics where the content-based feature is related to color histogram such as “black and white” in topic 11 or “night shots” in 15. We see improvements with different explanation as well (27: “motorcyclist racing”, 6: “straight road”) that must have also utilized certain semantical image content amplified in addition by biclustering. As an interesting example, Topic 53 “asymmetric stones” has a deterioration of 0.12 by visual similarity but an improvement of 0.18 by biclustering that sums up to +0.06 with a possible reason that biclustering removes segments belonging to the people in one of the sample images.

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop (2007)
2. Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K.: Cross-language retrieval with wikipedia. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2007)
3. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.* 5, 913–939 (2004)
4. Prasad, B.G., Biswas, K.K., Gupta, S.K.: Region-based image retrieval using integrated color, shape, and location index. *Comput. Vis. Image Underst.* 94(1-3), 193–233 (2004)
5. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.* 24(8), 1026–1038 (2002)
6. Lv, Q., Charikar, M., Li, K.: Image similarity search with compact data structures. In: *CIKM 2004: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 208–217. ACM Press, New York (2004)
7. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. *International Journal of Computer Vision* 59 (2004)
8. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 89–98 (2003)

Analysing an Approach to Information Retrieval of Visual Descriptions with IR-n, a System Based on Passages

Sergio Navarro, Fernando Llopis, Rafael Muñoz Guillena, and Elisa Noguera

Natural Language Processing and Information Systems Group,
University of Alicante, Spain

snavarro, llopis, rafael, elisa@dlsi.ua.es

<http://gplsi.dlsi.ua.es>

Abstract. This paper analyses an approach made to the development of a textual image retrieval system by the University of Alicante using IR-n, a text-based information retrieval (IR) system. With only a minimal amount of adaptations to the features of this task, our system has obtained precision results above the mean average of participants for ImageCLEF07 both for English (0.1604 vs 0.1388) and for Spanish (0.1482 vs 0.1450). For German, our results were below the mean (0.0991 vs 0.1331), which could be due to the fact that our system does not incorporate a splitter for the treatment of this agglutinative language. These results are obtained without the incorporation of image recovery dominion techniques. The error analysis shows us that there is still a considerable amount of work to do concerning text-based techniques in order to improve the system, but also shows that the key to successful participation in this task is to mix text and image resources.

1 Introduction

The number of documents that our information society produces is huge. We generate all kinds of documents: Plain-text documents, images, videos, source code, and many more. This quantity of documentation gives rise to a need for automatic techniques for accessing it.

Information retrieval (IR) systems are used to meet this challenge. The primary aim of such systems is to locate the documents in a document collection that are relevant for a users query.

The CLEF has organised a series of evaluation campaigns that aim to encourage the development of IR systems based on various different European languages. These projects take the form of annual competitions in which different IR systems compete against each other. A specific CLEF [1] area specializes in image retrieval - ImageCLEF [2] [3].

For our ImageCLEF task, we used IR-n [4]. This is an information retrieval system that uses statistical techniques and has yielded good results in plain-text-based tasks. The aim of using a system with these characteristics is to contrast

with regard to image recovery the results of a statistical system based only on text with other systems that also use NLP and/or image processing techniques.

This paper is structured as follows: First, it presents the main characteristics of the IR-n system, then it moves on to explain the process we used to evaluate the system and the training we carried out, and finally it describes our results and conclusions.

2 The IR-n System

In our approach we used IR-n - an information retrieval system based on passages. Passage-based IR systems treat each document as a set of passages, with each passage delimiting a portion of text or contiguous block of text. Unlike document-based systems, these systems can consider the proximity of words that appear in a document to each other in order to evaluate their relevance [4].

The IR-n passage-based IR system differs from other system in the same category with regard to the method proposed for defining the passages. It uses phrases as the unit for defining passages. This means that passages are defined by a number of consecutive phrases in a document [4].

This section describes the main characteristics of the IR-n system and details the techniques used for ImageCLEF 2007.

2.1 Resources: Stemmers and Stopword List

Stemmers and stopwords lists are used to determine which information in a document is to be used for retrieval. The stopwords list for each language contains words whose presence in a document is not considered important enough to determine that the document in question is relevant even if these words do appear in a query. Stemmers, on the other hand, are used to obtain the root of a word, thus eliminating any suffixes or prefixes, for indexing and search purposes. For a list of the stemmers and stopwords used by IR-n, see www.unine.ch/info/clef.

2.2 Weighting Models

Weighting models allow the quantification of the similarity between a text (a complete document or a passage in a document) and a query. Values are based on the terms that are shared by the text and query and on the discriminatory importance of each term. IR-n uses several weighting models. For this competition, we have used DFR [5] and OKAPI [6].

2.3 Query Expansion

Most IR systems use query expansion techniques [7] which are based on adding the most frequent terms contained in the most relevant documents to the original query. The IR-n architecture allows us to use query expansion based on either the most relevant passages or the most relevant documents. In previous studies, we obtained better results using the most relevant passages.

3 Training

IR-n is a parameterizable system, which means that it can be adapted in line with the concrete characteristics of the task at hand. The parameters for this configuration are the number of phrases that form a passage, the weighting model to use, the type of expansion, the number of documents/passages on which the expansion is based, and the average number of words per document. This section describes the training process that was carried out in order to obtain the best possible features for improving the performance of the system. The collections and resources are described first, and the next section describes specific experiments.

3.1 Data Collections

We participated in the following monolingual tasks within the framework of ImageCLEF 2007: English, German, and Spanish. The training for English and German was based on corpuses from previous years. Table 1 shows the characteristics of the language collections.

Table 1. Data Collections

Language	Collection	No. of documents	WDAvg	NrQ
English	St Andrews(ImageCLEF2004)	28.133	48	25
English	IAPR TC-12 (ImageCLEF2006)	20.000	40,29	60
German	IAPR TC-12 (ImageCLEF2006)	20.000	34,61	60

- **WDAvg**: is the average of words by document.
- **NrQ**: is the number of queries that are used in the experiments on each collection.

WDAvg is the average number of words per document and NrQ is the number of queries used in the experiments on each collection.

We did not have an existing corpus for Spanish, which is why the results obtained for English and German are used as a guide. As can be seen in Table 1, the difference between the corpus for 2004 and the corpus for 2006 is substantial, since the 2004 corpus contains a far greater amount of information (more documents and a larger average number of words per document). In addition, for the 2006 corpus only 70% of the documents contain all required information - 10% have no description, another 10% have only a location and date, and a further 10% have no notes. This means that chances of success when attempting to retrieve textual information are higher for the 2004 corpus. Another important difference between the 2004 and the 2006 corpus is that the textual information in the 2004 corpus takes the form of plain text, whereas the 2006 corpus uses a semi-structured format based on XML. This semi-structured format enables us to reduce the number of phrases per document to only the phrases in our fields of interest for the 2006 corpus. We therefore used the five text fields corresponding to the title (TITLE), description (DESCRIPTION), notes (NOTES), place

(LOCATION), and date of the photo (DATES) as input for IR-n. In the case of the 2004 corpus, we used entire documents as the input for IR-n. It is important to remember that this kind of input will result in a greater number of phrases per document than the input for the 2006 corpus, since the documents in the 2004 corpus include more fields of information than those in the 2006 corpus (record ID, short title, long title, location, description, date, photographer, categories, and notes) and each field is entered in a new line (meaning that there is at least one new phrase per field). The queries also have a semi-structured format, but only the English queries in ImageCLEF 2006 and ImageCLEF 2004 contain narrative (NARR) that accompanies the query (TITLE).

3.2 Experiments

The experiment phase aims to establish the optimum value of the input parameters for each collection. Below is a description of the input parameters of the system:

- **Size of the Passage (sp):** Number of phrases in a passage.
- **Weight model (wm):** We used two weighting models : **OKAPI** and **DFR**.
- **Opaki parameters:** k_1 , b and avg_{ld} (k_3 is fixed as 1000).
- **DFR parameters:** c and avg_{ld} .
- **Query expansion parameters:** If **exp** has value 1, this denotes we use relevance feedback based on passages. But, if **exp** has value 2, the relevance feedback is based on documents. Moreover, **num** denote the number of passages or documents that the expansion will use, and **term** indicates the k terms extracted from the best ranked passages or documents from the original query
- **Evaluation Measure:** MAP or Mean Average Precision (**avgP**) is the evaluation measure used in order to evaluate the experiments.

English. As we can see from Table 2 which shows the results of the experiments on the 2004 corpus query expansion in general (and, in particular, passage-based expansion) improves the results for both weight models. For this corpus, the best results are obtained with a passage size of five phrases. The results for smaller and larger passage sizes show that the precision is less. It is important to point out that better results would be obtained if using documents with a larger average length (in bytes) than in this corpus.

We can see that the level of precision is considerably reduced for the 2006 corpus. This is justified by the fact that the corpus has only a small amount of textual information compared with the 2004 corpus, reflected in Table 1. The fact that the 2006 corpus has 30% of documents that are incomplete also plays a part. In addition, we observed that the best results are obtained with DFR and techniques based on query expansion with a passage size of three phrases. It is important to point out that the optimum passage size for this corpus is three phrases.

German. As shown in Table 4, the results for German are lower. This could be due to a combination of causes: On the one hand, IR-n does not have a mechanism for treating compound words (which are very common in German).

Table 2. English 2004 - Best Results

sp	wm	c	avgl	d	k1	b	exp	num	term	avgP
3	DFR	3	350							0.4406
4	DFR	4	350							0.4573
5	DFR	4	1600							0.4752
3	OKAPI		90	2	0.6					0.4460
4	OKAPI		350	4	0.8					0.4572
5	OKAPI		50	1	0.2					0.4752
3	DFR	2	350			2	10	5		0.4700
4	DFR	4	1200			2	10	5		0.4843
5	DFR	9.5	1700			1	10	5		0.5128
3	OKAPI		90	3	0.4	2	10	5		0.4768
4	OKAPI		350	3	0.2	1	5	5		0.4852
5	OKAPI		350	3	0.2	1	5	5		0.5086

Table 3. English 2006 - Best Results

sp	wm	c	avgl	d	k1	b	exp	num	term	avgP
2	DFR	5.5	85							0.1926
2	OKAPI		90	4	0.8					0.1799
2	OKAPI		1900	4	0.8					0.1800
3	DFR	8	85			2	5	5		0.2059
5	OKAPI		90	2	0.8	1	10	10		0.1992

Table 4. German 2006 - Best Results

sp	wm	c	avgl	d	k1	b	exp	num	term	avgP
3	DFR	2	90							0.1487
3	OKAPI		85	2	0.8					0.1492
3	DFR	2	85			1	5	5		0.1742
3	OKAPI		85	2	0.8	2	5	5		0.1857

On the other hand, the German queries do not contain narrative, which reduces the chance of success. For German, the configuration that yields the best results is a passage size of three phrases, query expansion on the basis of documents, and OKAPI as the weighting model.

Experiments Summary. An important finding of these experiments is that the most suitable passage size is not as long as the size of the documents but seems to be proportional to the number of phrases in the documents. Another important conclusion is that passage-based expansion improves results for the corpus with more phrases and with a variety of types of information per document (that is, the 2004 corpus). However, for a corpus with fewer phrases and more uniformity with regard to the type of information (that is, the 2006 corpus), the best results are obtained with document-based expansion.

Table 5. Comparison of Results

Competition	avgP IR-n	avgP ImageCLEF	avgP Best
ImageCLEF04 English	0.5128	0.4155	0.58
ImageCLEF06 English	0.2059	0.152	0.385
ImageCLEF06 German	0.1857	0.121	0.311

Table 5 compares the best results of our training phase with the results that were obtained by participants in the various ImageCLEFs.

For our participation in ImageCLEF07, we used the same configuration as the best one used for the ImageCLEF06 corpus. This is because the documents in the ImageCLEF07 corpus have the same fields (apart from one the description field) as the documents in the ImageCLEF06 corpus.

4 Results at ImageCLEFPhoto-2007

Table 6 shows the configurations used for ImageCLEF 2007 and a comparison of the results obtained along with an average for all participants by language for monolingual tasks. For English and German, we used the configurations that obtained the best results during the training phase using the ImageCLEF06 corpus. For Spanish, we used the same configuration as for English. By analysing the mistakes made by the system and comparing the system with other ImageCLEF projects, we observed three main causes of error: Unlike other systems, our system has no image content retrieval. It also has no mechanisms for focusing retrieval on geographical zones or animal groups, and the local expansion procedure often includes incorrect terms whilst ignoring important terms that might otherwise improve the results.

Table 6. ImageCLEFPhoto 2007 official average results - monolingual tasks

lang	sp	wm	c	avgl	d	k1	b	exp	num	term	avgP	ImgCLEF07
Eng	3	DFR	8	85				2	5	5	0.1604	0.1388
	3	DFR	8	85				0	0	0	0.1453	
Spa	3	DFR	8	85				2	5	5	0.1482	0.1450
	3	DFR	8	85				0	0	0	0.1367	
Ger	3	OKAPI		85	2	0.8	2	5	5	5	0.0991	0.1331
	3	OKAPI		85	2	0.8	0	0	0	0	0.0911	

Table 7 shows that the most common error for all queries is an insufficiently good selection of terms for the expansion and that only 15% of queries are failed by the IR-n system because there is no way other than image content analysis to suitably retrieve the information required by the query.

Table 7. Analysis of main errors

Main Error	Queries
Insufficient expansion with other problems	47%
No geographical or biological specialization with other problems	41%
No image content with other problems	38%
No geographical or biological specialization only	20%
No image content analysis only	15%
Insufficient expansion only	13%

5 Conclusion and Future Work

For our first ImageCLEF participation, we used a text-based IR system. We made a minimum number of adaptations to the image recovery dominion. We would like to point out that precision results are above average for both English and Spanish, which leads us to conclude that we have here a very good basis from which to work towards obtaining better results.

Analysing the results of training with the corpuses from previous years allows us to measure the extent to which the incompleteness of the corpuses used and the absence of document classification (like in the 2004 corpus) result in a reduction in the precision values. Moreover, the reduction in the length of the documents has a direct effect on the passage size. In fact, an important conclusion of our experiments is that the most suitable passage size is not as long as the size of the documents. This confirms the benefits of using a passage-based IR system, since although the size of the documents is small, using a smaller passage size enables higher relevance when evaluating documents in which the terms in the query appear in close proximity.

Although Table 7 shows that adding improvements based on textual information will yield a significant improvement in the overall results for our system rather than an improvement based on image content, other projects in this ImageCLEF session [8,9,10], and Table 7 itself, demonstrate that a multimedia approximation along with a suitable mixing procedure constitutes the key to successful participation in this task.

In order to achieve the continuing improvement of the system, we shall attempt to add natural language processing techniques to the local query expansion system, thereby improving the quality of the expansion terms. Moreover, we shall explore the extraction of features from images.

Acknowledgements

This research has been partially funded by the Spanish Government within the framework of the TEXT-MESS (TIN-2006-15265-C06-01) project and by European Union (EU) within the framework of the QALL-ME project (FP6-IST-033860).

References

1. <http://www.clef-campaign.org>
2. <http://ir.shef.ac.uk/imageclef>
3. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF-photo 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
4. Llopis, F.: IR-n: Un Sistema de Recuperacin de Informacin Basado en Pasajes. PhD thesis, University of Alicante (2003)
5. 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)
6. J., S.: 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. Chen, A., Gey, F.C.: Combining Query Translation and Document Translation in Cross-Language Retrieval. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237. Springer, Heidelberg (2004)
8. Gao, S., Chevallet, J.P., Le, T.H.D., Pham, T.T., Lim, J.H.: Ipal at imageclef 2007 mixing features, models and knowledge. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
9. Wilhelm, T., Krsten, J., Eibl, M.: Experiments for the imageclef 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
10. Villena-Romn, J., Lana-Serrano, S., Martnez-Fernndez, J.L., Gonzlez-Cristbal, J.C.: Miracle at imageclefphoto 2007:evaluation of merging strategies for multilingual and multimedia information retrieval. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)

DCU and UTA at ImageCLEFPhoto 2007

Anni Järvelin¹, Peter Wilkins², Tomasz Adamek², Eija Airio¹,
Gareth J.F. Jones², Alan F. Smeaton², and Eero Sormunen¹

¹ University of Tampere (UTA), Finland
`Anni.Jarvelin@uta.fi`

² Dublin City University (DCU), Ireland
`pwilkins@computing.dcu.ie`

Abstract. Dublin City University (DCU) and University of Tampere (UTA) participated in ImageCLEF 2007 photographic retrieval task with several monolingual and bilingual runs. The approach was language independent with text retrieval utilizing fuzzy *s*-gram query translation and combined with visual retrieval. Data fusion was achieved through unsupervised query-time weight generation approaches. The baseline was a combination of dictionary-based query translation and visual retrieval, which achieved the best result. The best mixed modality runs using fuzzy *s*-gram translation reached on average around 83% of the baselines' performance. This approach was much closer at the early precision levels of P@10 and P@20. This suggests that our language independent approach could be a cheap alternative for cross-lingual image retrieval. Both sets of results further emphasize the merit in our query-time weight generation schemes for data fusion, with the fused runs exhibiting marked performance increases over single modalities without the use of prior training data.

1 Introduction

Retrieving images by their associated text is a common approach in image retrieval [1]. When cross-language image retrieval is considered, this approach requires language dependent linguistic resources for query translation. Machine-readable dictionaries, machine translation tools or corpus-based translation tools are expensive and not available for all the language pairs. However, there are alternative approaches which may be used to compensate linguistic tools, for example the fuzzy string matching technique *n*-gram matching and its generalization *s*-gram matching. These techniques have previously been used for translation of query words missing from dictionaries [2][3], but only rarely for the whole query translation [4][5].

In earlier ImageCLEF campaigns combined text and visual retrieval approaches have performed better than text or image retrieval alone. In this year's campaign, text retrieval faced a new challenge of retrieval of lightly annotated photographs [1]. A negative impact on the performance of the text retrieval techniques was to be expected and therefore successful fusion of text and visual

features became even more important. We tested a language independent image retrieval approach, where *s*-gram-based fuzzy query translations were fused with visual retrieval. We explored data fusion techniques with query-time coefficient generation for retrieval expert combination. We experimented primarily with altering the stages at which we fuse various experts together. For instance we experimented with fusing all the visual experts into a single expert, then fusing with text, as opposed to treating all experts equally. To have a strong baseline, the performance of the language independent approach was compared to a combination of dictionary-based query translation and visual retrieval.

To study the effect of the source and target languages on the quality of the fuzzy translation we selected six language pairs where source/target languages were related to each other at different levels. The Scandinavian languages Danish, Norwegian and Swedish were translated into German, to which they are quite closely related to. French was the source language that was closest to English. German and English are not very closely related and translation between them was done both ways. (See [4] for a matrix for similarities between English, French, German and Swedish) A total of 138 runs were submitted. Reporting the results for all of these would be impractical and therefore only the results for the most interesting runs are presented here.

2 Background

***s*-gram-based query translation.** The *s*-gram matching is a fuzzy string matching technique that measures similarity between text strings. The text strings to be compared are decomposed into substrings (*s*-grams) and the similarity is calculated as the overlap of their common substrings. *s*-gram matching is a generalization of *n*-gram technique, commonly used for matching cognates in Cross Language Information Retrieval (CLIR). While the *n*-gram substrings consist of adjacent characters of the original strings, skipping some characters is allowed when forming the *s*-grams. In classified *s*-gram matching [6], different types of *s*-grams are formed by skipping different number of characters. The *s*-grams are then classified into sets based on the number of characters skipped and only the *s*-grams belonging to the same set are compared to each other when calculating the similarity. Character Combination Index (CCI) indicates the set of all the *s*-gram types to be formed from a string. CCI $\{\{0\}, \{1, 2\}\}$ for example means that three types of *s*-grams are formed and classified into two sets: one set of conventional *n*-grams formed of adjacent characters ($\{0\}$) and one of *s*-grams formed both by skipping one and two characters ($\{1, 2\}$). For an extensive description of the *s*-gram matching technique, see [6][7].

Proper names are very common query terms when searching from image databases [8] and are typically not covered by translation dictionaries and thus remain untranslatable in queries. Proper names in related languages are often spelling variants of each other and can thus be translated using approximate string matching.

Visual Retrieval. To facilitate visual retrieval we made use of six ‘low-level’ global visual experts. Our visual features are MPEG7 features and were extracted using the *aceToolBox*, developed as part DCU’s collaboration in the *aceMedia* project [9]. These six features included: Colour Layout, Colour Structure, Colour Moments, Scalable Colour, Edge Histogram and Homogenous Texture. Further details on these descriptors can be found in [10] and [11]. Distance metrics for each of these features were implementations of those specified in the MPEG7 specification [11].

Query-Time Fusion. The combination of retrieval experts for a given information need can be expressed as a *data fusion* problem [12]. Given that for any given information need different retrieval experts perform differently, we require some form of weighting scheme in order to combine experts. Typical approaches to weight generation include the use of global query-independent weights or query-class expert weights learnt through experimentation on a training collection to name a few.

Our approach to weight generation differs in that it is a query-dependant weighting scheme for expert combination which requires no training data [13]. This work was based upon an observation, that if we were to plot the normalized scores of an expert, against that of scores of other experts used for a particular query, that the expert who’s scores exhibited the greatest initial change in scores correlated with that expert being the best performer for that query. Examples of this observation can be seen in [13] and our ImageCLEF workshop paper [14]. This approach is not giving us any absolute indication of expert performance, which other approaches to examining score distributions attempt to provide, an excellent review of which is given by Robertson [15]. We would note that this observation is not universal, and we can identify failure cases where this observation will not occur. If we assume though that in a majority of queries this observation will hold, then we can employ techniques that leverage this approach to create query-time expert coefficients for data fusion. Our main technique involves measuring the change in scores for a retrieval expert over a top subset of its results, versus the change in scores over a larger sample of that experts scores. The expert which undergoes the greatest initial change in score is assigned a greater weight. A complete explanation of this process can be found in [13].

3 Resources, Methods and Runs

Text Retrieval and Indexing. We utilized the Lemur toolkit (Indri engine) [16] for indexing and retrieval. Indri combines language modeling to inference nets and allows structured queries to be evaluated using language modeling estimates. The word tokenization rules used in indexing included converting punctuation marks into spaces and capitals were converted into lower case. Strings separated by the space character were tokenized into individual words. For the dictionary-based translation, lemmatized English and German indices were created. The image annotation text was lemmatized, words not recognized by the lemmatizers were indexed as such. Compound words were split and both the

original compound and the constituents were indexed. For the *s*-gram-based translation we used inflected indices, where the words were stored in the inflected word forms in which they appeared in the image annotations. Stop words were removed.

Topics were processed as the indices. For the *s*-gram-based translation stop words were removed from the queries and the remaining words were translated into the target language with *s*-gram matching. The CCI was set to be $\{\{0\}, \{1, 2\}\}$, and the Jaccard coefficient [7] was used for calculating the *s*-gram proximity. Three best matching index words were selected as translations for each topic word. A similarity threshold value of 0.3 was set to discard bad translations, only the index words exceeding this threshold with respect to a source word were accepted as translations. As a consequence some of the query words could not be translated. Queries were structured utilizing a synonym operator where target words derived from the same source word were grouped into the same synonym group (*the Pirkola method*, [17]). Indris Pseudo Relevance Feedback (PRF) was used with 10 keys from the 10 highest ranked documents in the original result list.

For the dictionary-based query translation, the UTACLIR query translation tool was used. UTACLIR was originally developed for the CLEF 2000 and 2001 campaigns [2]. It utilizes external language resources, such as translation dictionaries, stemmers and lemmatizers. Topic words were lemmatized, stop words removed and finally the non-stop words translated. Next, untranslatable compound words were split and the constituents were translated. Translation equivalents were normalized utilizing a target language lemmatizer. Untranslatable words were matched against the database index using the *s*-gram matching. Queries were structured with the synonym operator. A morphological analyzer for French was not available and therefore the French topics were analyzed manually. This might have resulted in a slightly better quality of lemmatization than automatic analysis, even though we strived for comparable quality. We used PRF with 20 expansion keys from the 15 top ranked documents.

Data Fusion. The query-time data fusion approach specified in Section 2 describes our basic approach to expert combination. However, one set of parameters that was not specified was the order in which experts will be combined. This is the focus of our experimental work in this section.

One commonality between all the combination approaches we try in this work is the fusion of the low-level visual experts. For each query image we fuse the six low-level visual experts into a single result for each image, where the combination of these is using the aforementioned technique. Therefore for each query, the visual component was then represented by three result sets, one for each query image. Additionally for a subset of our runs we introduce a seventh visual expert, the FIRE baseline [18]. In cases where FIRE was used, because it was a single result for the three visual query images, we first combined our MPEG7 visual features into a single result for each image, then these combined into an overall

image result, which was then combined with the FIRE baseline. There are four variants that we tried in our combination work, which are:

- dyn-equal: Query-time weighting method used, text and individual image results combined at the same level (i.e. we have three image results and one text result which is to be combined).
- dyn-top: As above, except the results for each query image were fused into a single image result, which was then combined with the text result (i.e. image results combined into a single result, which is then combined with the single text result).
- stat-eventop: Query-time weighting to produce single image result list, image and text fused together with equal static weighting (0.5 coefficient).
- stat-imgHigh: As above, except with the image result assigned a static weight of 0.8 and text a static weight of 0.2.

Additionally, any of our runs which ended in ‘fire’ incorporated the FIRE baseline into the set of visual experts used for combination.

4 Results

Our tables of results are organized as follows. Table 1 presents our baseline runs, including monolingual text-only, visual-only and baseline fusion results mixing these two types. Table 2 presents our central cross-lingual results with mixed modalities. In all tables where data fusion is utilized, we present only the best performing data fusion approach. Except where noted, all visual results used in data fusion presented here incorporated the FIRE baseline as visual data which included the FIRE baseline with our global MPEG7 features consistently outperformed global MPEG7 by themselves.

For the monolingual runs in Table 1, the runs where morphological analysis (dict) was used performed slightly better than the *s*-gram runs. The difference is small for the English runs. For German runs the difference is greater, which is understandable as German has a more complex inflectional morphology than

Table 1. ImageCLEF Baseline Results

Language Pair	Modality	Text	Fusion	FB	MAP	P@10	P@20
EN-EN	Text	dict	n/a	no	0.1305	0.1550	0.1408
EN-EN	Text	<i>s</i> -gram	n/a	yes	0.1245	0.1133	0.1242
DE-DE	Text	dict	n/a	yes	0.1269	0.1717	0.1533
DE-DE	Text	<i>s</i> -gram	n/a	yes	0.1067	0.1233	0.1125
MPEG7 With FIRE	Visual	na	dyn-equal	no	0.1340	0.3600	0.2658
MPEG7 Without FIRE	Visual	na	dyn-equal	no	0.1000	0.2700	0.1958
EN-EN	Mixed	dict	dyn-equal	yes	0.1951	0.3967	0.3150
EN-EN	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1833	0.3833	0.3092
DE-DE	Mixed	dict	dyn-equal	yes	0.1940	0.4033	0.3300
DE-DE	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1628	0.3350	0.2792

English. Our text and visual retrieval techniques were almost equal, which is notable in the context of earlier years' ImageCLEF results. Our best visual-only run performed well being the second best visual approach in terms of Mean Average Precision (MAP). Its MAP value 0.1340 is comparable to our best monolingual English text run scoring 0.1305. We believe that the comparative low performance of the text expert (when compared to the dominance of text in previous years) was due to the reduced length of the annotations for 2007. Table 1 also presents our fused monolingual text and visual retrieval runs, which performed clearly better than any of the text or visual runs alone. Fusion of these modalities produced consistent improvements in MAP of between 65% and 67%.

From a data fusion perspective, no single approach of the four we tried consistently performed the best. Whilst our results presented here show the "dyn-equal" fusion as being superior, this is because it was the only fusion type attempted with visual data which incorporated the FIRE baseline. For runs where FIRE was not used, there best performing fusion type varied depending on the text type (dictionary or sgram) or language pair used. In a majority of cases all fusion types performed similarly, as such deeper investigation with significance testing will be required in order to infer any meaningful interpretations. However, we can conclude that as all four fusion types made use of our query-time weight generation method at some level, that this technique is capable of producing weights which lead to performance improvements when combining results. What is unknown is how far from the optimal query-dependant combination performance we achieved, and that will be one of the ultimate measures of the success of this approach.

Table 2 presents our central cross-lingual results. Dictionary-based query translation was the best query translation approach. The best mixed modality runs using the *s*-gram-based query translation nevertheless reached on average around 84% of the MAP of the best mixed modality runs using dictionary-based translation. The difference between the approaches further decreased when the early precision values of P@10 and P@20 were considered. The best *s*-gram runs

Table 2. ImageCLEF CLIR Fusion Results

Language Pair	Modality	Text	Fusion	FB	MAP	P@10	P@20
FR-EN	Mixed	dict	dyn-equal	yes	0.1819	0.3583	0.2967
FR-EN	Mixed	<i>s</i> -gram	dyn-equal	no	0.1468	0.3483	0.2667
DE-EN	Mixed	dict	dyn-equal	yes	0.1910	0.3483	0.3042
DE-EN	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1468	0.3233	0.2533
DA-DE	Mixed	dict	dyn-equal	yes	0.1730	0.3467	0.2942
DA-DE	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1572	0.3350	0.2717
NO-DE	Mixed	dict	dyn-equal	yes	0.1667	0.3517	0.2700
NO-DE	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1536	0.3167	0.2667
SV-DE	Mixed	dict	dyn-equal	yes	0.1788	0.3817	0.2942
SV-DE	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1472	0.3050	0.2500
EN-DE	Mixed	dict	dyn-equal	yes	0.1828	0.3633	0.3008
EN-DE	Mixed	<i>s</i> -gram	dyn-equal	yes	0.1446	0.3350	0.2667

reached on average around 91% of the best dictionary-based runs performance at P@10 and around 89% at P@20. If the high ranks of the result list are considered to be important from the user perspective, the *s*-gram translation could be seen as almost equal with the dictionary-based translation in mixed modality runs. These results varied depending on the language pair. *s*-gram-based and dictionary-based translation performed similarly for the closely related language pairs, while the differences were greater for the more distant language pairs. The *s*-gram translation reached its best results with Norwegian and Danish topics and German annotations - over 90% of the dictionary translation's MAP, and the worst ones between German and English - less than 80% of the dictionary translation's MAP.

5 Conclusions

In this paper we have presented the joint DCU and UTA ImageCLEF 2007 Photo results. In our work we experimented with two major variables, that of cross-lingual text retrieval utilizing minimal translation resources, and query-time weight generation for expert combination. Our results are encouraging and support further investigation into both approaches. Further work is now required to conduct a more thorough analysis of contributing factors to performance emphasized by each approach. Of particular interest will be the degree to which each of these approaches introduced new information, or re-ordered existing information presented by the systems. For instance, we do not know yet if the *s*-gram retrieval found relevant documents that were missed by the dictionary based approach. Likewise for data fusion, we do not know yet if we promoted into the final result set relevant results which were only present in some and not all of the experts used.

Acknowledgments

We are grateful to the AceMedia project (FP6-001765) which provided the image analysis toolkit. Research leading to this paper was supported by the European Commission under contract FP6-027026 (K-Space). The work of the first author is funded by Tampere Graduate School of Information Science and Engineering (TISE).

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF-photo 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
2. Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., Järvelin, K.: Utaclir @ CLEF 2001 - effects of compound splitting and n-gram techniques. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 118–136. Springer, Heidelberg (2002)

3. Hiemstra, D., Kraaij, W.: Twenty-One at TREC7: Ad-hoc and cross-language track. In: TREC, pp. 174–185 (1998)
4. Mcnamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. *Information Retrieval* 7(1-2), 73–97 (2004)
5. Järvelin, A., Järvelin, A., Järvelin, K.: s-grams: Defining generalized n-grams for information retrieval. *Information Processing and Management* 43(4), 1005–1019 (2007)
6. Pirkola, A., Keskustalo, H., Leppänen, E., Käsälä, A.P., Järvelin, K.: Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants. *Information Research* 7(2) (2002)
7. Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., Järvelin, K.: Non-adjacent digrams improve matching of cross-lingual spelling variants. In: SPIRE, pp. 252–265 (2003)
8. Markkula, M., Sormunen, E.: End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval* 1(4), 259–285 (2000)
9. AceMedia: The AceMedia Project, <http://www.acemedia.org>
10. O'Connor, N., Cooke, E., le Borgne, H., Blighe, M., Adamek, T.: The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In: 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (2005)
11. Manjunath, B., Salembier, P., Sikora, T. (eds.): Introduction to MPEG-7: Multimedia Content Description Language. Wiley, Chichester (2002)
12. Belkin, N.J., Kantor, P., Fox, E.A., Shaw, J.A.: Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management* 31(3), 431–448 (1995)
13. Wilkins, P., Ferguson, P., Smeaton, A.F.: Using score distributions for querytime fusion in multimedia retrieval. In: MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (2006)
14. Jarvelin, A., Wilkins, P., Adamek, T., Airio, E., Jones, G., Smeaton, A.F., Sormunen, E.: DCU and UTA at ImageCLEFPhoto 2007. In: ImageCLEF 2007 - The CLEF Cross Language Image Retrieval Track Workshop (2007)
15. Robertson, S.: On score distributions and relevance. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 40–51. Springer, Heidelberg (2007)
16. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language-model based search engine for complex queries (extended version) (2005-02-14) (2005)
17. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: SIGIR 1998: Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55–63 (1998)
18. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in 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. 652–661. Springer, Heidelberg (2006)

Cross-Language and Cross-Media Image Retrieval: An Empirical Study at ImageCLEF2007

Steven C.H. Hoi

School of Computer Engineering
Nanyang Technological University
Singapore 639798
{chhoi}@ntu.edu.sg

Abstract. This paper summarizes our empirical study of cross-language and cross-media image retrieval at the CLEF image retrieval track (ImageCLEF2007). In this year, we participated in the ImageCLEF photo retrieval task, in which the goal of the retrieval task is to search natural photos by some query with both textual and visual information. In this paper, we study the empirical evaluations of our solutions for the image retrieval tasks in three aspects. First of all, we study the application of language models and smoothing strategies for text-based image retrieval, particularly addressing the short text query issue. Secondly, we study the cross-media image retrieval problem using some simple combination strategy. Lastly, we study the cross-language image retrieval problem between English and Chinese. Finally, we summarize our empirical experiences and indicate some future directions.

1 Introduction

Digital image retrieval has attracted a surge of research interests in recent years. Most existing Web search engines usually search images by text only. They have yet to solve the retrieval tasks very effectively due to unreliable text information. Until now, general image retrieval is still a challenging research task. In this paper, we study the methodology of cross-language and cross-media retrieval techniques to attack some open challenges at ImageCLEF.

In this participation, we offer major contributions in three aspects. Firstly, we study an empirical evaluation of language models and smoothing strategies for cross-language image retrieval. Secondly, we conduct an evaluation of cross-media image retrieval, i.e., combining text and visual contents for image retrieval. The last contribution is the empirical evaluation of a methodology for bilingual image retrieval spanning English and Chinese sources.

The rest of this paper is organized as follows. Section 2 reviews some methodology of the TF-IDF retrieval model and the language model for information retrieval. Section 3 presents our implementation for this participation, and outlines our empirical study on cross-language and cross-media image retrieval. Section 4 set out our conclusions.

2 Review of Language Models and Smoothing Techniques

In our approaches, we have conducted an extensive set of experiments to evaluate the performance of state-of-the-art language models and smoothing techniques with applications to text-based image retrieval tasks. Specifically, two retrieval models are studied: (1) the TF-IDF (Term Frequency-Inverse Document Frequency) model, and (2) the KL-divergence language model. Three smoothing strategies [1] are evaluated: (1) the Jelinek-Mercer (JM) method, (2) Bayesian smoothing with Dirichlet priors (DIR), and (3) Absolute discounting (ABS).

2.1 TF-IDF Retrieval Model

The TF-IDF retrieval model is a well-known method for text-based retrieval [2]. In general, a document and a query can be represented as a term frequency vector $\mathbf{d} = (x_1, x_2, \dots, x_n)$ and $\mathbf{q} = (y_1, y_2, \dots, y_n)$ respectively, where n is the number of total terms, x_i and y_i are the frequency (counts) of term t_i in the document vector $\mathbf{d}$ and query vector $\mathbf{q}$, respectively. In a retrieval task, given a document collection $\mathcal{C}$, the IDF of a term t is defined by $\log(N/n_t)$, where N is the total number of documents in $\mathcal{C}$, and n_t is the number of documents that contain the term t . For the TF-IDF representation, all terms in the query and documents vectors are weighted by the TF-IDF weighting formula, i.e., $\mathbf{d}' = (tf_d(x_1)idf(t_1), tf_d(x_2)idf(t_2), \dots, tf_d(x_n)idf(t_n))$ and $\mathbf{q}' = (tf_q(y_1)idf(t_1), tf_q(y_2)idf(t_2), \dots, tf_q(y_n)idf(t_n))$. For a simple TF-IDF retrieval model, one simply takes $tf_d(x_i) = x_i$. One can also define some other heuristic formula for the TF function. For example, the Okapi retrieval approach is a special case of TF-IDF model by defining the document TF formula [3] as: $tf_d(x) = \frac{k_1 x}{x + k_1(1 - b + b \frac{l_d}{l_C})}$, where k_1 and b are two parameters for the document TF function, l_d and l_C are the lengths of the given document and collection, respectively. Similarly, a query TF function can be defined with parameters k_1 and b as well as l_q representing the average length of queries. In TF-IDF retrieval models, cosine similarity is often adopted as similarity measure.

2.2 Language Modeling for Information Retrieval

Language model, or the statistical language model, employs a probabilistic mechanism to generate text. The earliest serious approach for a statistical language model may be tracked to Claude Shannon [4]. To apply his newly founded information theory to human language applications, Shannon evaluated how well simple n -gram models did at predicting or compressing natural text. In the past, there has been considerable attention paid to using the language modeling techniques for text document retrieval and natural language processing tasks [5].

The KL-Divergence Measure. Given two probability mass functions $p(x)$ and $q(x)$, $D(p||q)$, the Kullback-Leibler (KL) divergence (or relative entropy) between p and q is defined as $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$. One can show that $D(p||q)$ is always non-negative and is zero if and only if $p = q$. Even though it

is not a true distance between distributions (because it is not symmetric and does not satisfy the triangle inequality), it is often still useful to think of the KL-divergence as a “distance” between distributions [6].

The KL-Divergence Based Retrieval Model. In the language modeling approach, we assume a query q is generated by a generative model $p(q|\theta_Q)$, where θ_Q denotes the parameters of the query unigram language model. Similarly, we assume a document d is generated by a generative model $p(q|\theta_D)$, where θ_Q denotes the parameters of the document unigram language model. Let $\hat{\theta}_Q$ and $\hat{\theta}_D$ be the estimated query and document models, respectively. The relevance of d with respect to q can be measured by the negative KL-divergence function [5]:

$$-D(\hat{\theta}_Q||\hat{\theta}_D) = \sum_w p(w|\hat{\theta}_Q) \log p(w|\hat{\theta}_D) + (-\sum_w p(w|\hat{\theta}_Q) \log p(w|\hat{\theta}_Q)) \quad (1)$$

In the above formula, the second term on the right-hand side of the formula is a query-dependent constant, i.e., the entropy of the query model $\hat{\theta}_Q$. It can be ignored for the ranking purpose. In general, we consider the smoothing scheme for the estimated document model as follows:

$$p(w|\hat{\theta}_D) = \begin{cases} p_s(w|d) & \text{if word } w \text{ is present} \\ \alpha_d p(w|\mathcal{C}) & \text{otherwise} \end{cases} \quad (2)$$

where $p_s(w|d)$ is the smoothed probability of a word present in the document, $p(w|\mathcal{C})$ is the collection language model, and α_d is a coefficient controlling the probability mass assigned to unseen words, so that all probabilities sum to one [5]. We discuss several smoothing techniques in detail below.

2.3 Three Smoothing Techniques

In the context of language modeling, the term “smoothing” can be defined as the adjustment of the maximum likelihood estimator of a language model so that it will be more accurate [1]. As the maximum likelihood estimator often underestimates the probabilities of unseen words in the given document, it is important to employ smoothing methods that usually discount the probabilities of the words seen in the text and assign the extra probability mass to the unseen words according to some model [1]. Specifically, three representative smoothing methods are used in our scheme:

Jelinek-Mercer (JM) smoothing: a linear interpolation of the maximum likelihood model with the collection model, using a coefficient λ to control the influence: $p_\lambda(\omega|d) = (1 - \lambda)p_{ml}(\omega|d) + \lambda p(\omega|\mathcal{C})$, which is a simple mixture model [7].

Bayesian smoothing with Dirichlet Priors (DIR): the model is represented as: $p_\mu(\omega|d) = \frac{c(\omega;d) + \mu p(\omega|\mathcal{C})}{\sum_w c(\omega;d) + \mu}$, where μ in the is a DIR parameter that is estimated empirically from training sets [1].

Absolute discounting Smoothing (ABS): the model is represented as: $p_\delta(\omega|d) = \frac{\max(c(\omega;d) - \delta, 0)}{\sum_w c(\omega;d)} + \sigma p(\omega|\mathcal{C})$, where $\delta \in [0, 1]$ is a discount constant and

$\sigma = \delta|d|_{\mu}/|d|$, so that all probabilities sum to one. Here $|d|_{\mu}$ is the number of unique terms in document d , and $|d|$ is the total count of words in the document, i.e., $|d| = \sum_{\omega} c(\omega; d)$.

3 Cross-Language and Cross-Media Image Retrieval

The goal of the photographic retrieval task is to find as many relevant images as possible from an image collection given a multilingual statement describing a user information need. This task intends to simulate the text-based retrieval from photographs with multilingual captions, meanwhile queries for content-based image retrieval will also be offered. In this section, we study techniques to address several open challenges in this retrieval task, including (1) short text query problem, (2) cross-media image retrieval, and (3) cross-language retrieval. In the following section, we first describe the experimental testbed and setup at the ImageCLEF 2007, in which we have participated in the photo retrieval task. We will then conduct the empirical evaluations to address the above challenges and summarize our empirical experiences.

3.1 Experimental Testbed and Setup

The experimental testbed contains 20,000 color photographs with semi-structured captions in English, German and Spanish. For performance evaluations, there are 60 queries, each of them describes the user's information needs by short text in a range of languages including English, Italian, Spanish, French, German, Chinese, Japanese and Russian, and sample images.

For the photographic retrieval task, we have studied the query tasks in English and Chinese (simplified). Both text and visual information are used in our experiments. To evaluate the language models correctly, we employ the *Lemur* toolkit¹. A standard list of stopwords provided by the Lemur toolkit is used in the parsing step.

To evaluate the influence on the performance of using the different schemes, we have evaluated the methods by trying a variety of different configurations in order to examine every aspects of the solutions. In particular, three groups of performance evaluations will be studied in the subsequent parts.

3.2 Evaluation of Language Models and Smoothing Techniques

In our experiments, we study several retrieval methods by language models with different smoothing techniques for the text-based image retrieval tasks. Table 1 shows the results of a number of our submissions with respect to the text based retrieval approaches by Language Models. The listed methods are ranked by the MAP (mean average precision) score. From the results, we can observe that the best approach is the "Eng-kl-dir-fb2" solution, which is based on the KL-divergence language model with the Dirichlet priors smoothing technique. We

¹ <http://www.lemurproject.org/>.

Table 1. Evaluation of language models for text-based image retrieval tasks

Run ID	Method	Query	Source	Modality	RunType	QE/RF	MAP	P10	REL_RET	REL
Eng-kl-dir-fb2	KL-DIR	English	English	TEXT	AUTO	FB	0.1660	0.2217	1827	3416
Eng-kl-jm-fb1	KL-JM	English	English	TEXT	AUTO	FB	0.1641	0.2017	1788	3416
Eng-tf-idf-fb3	TF-IDF	English	English	TEXT	AUTO	FB	0.1641	0.2150	1955	3416
Eng-kl-jm-fb2	KL-JM	English	English	TEXT	AUTO	FB	0.1640	0.2033	1870	3416
Eng-kl-abs-fb2	KL-ABS	English	English	TEXT	AUTO	FB	0.1635	0.2017	1757	3416
Eng-okapi-fb2	OKAPI	English	English	TEXT	AUTO	FB	0.1612	0.2333	1674	3416
Eng-kl-abs-fb1	KL-ABS	English	English	TEXT	AUTO	FB	0.1611	0.1950	1700	3416
Eng-kl-dir-fb1	KL-DIR	English	English	TEXT	AUTO	FB	0.1603	0.2117	1682	3416
Eng-kl-abs-fb3	KL-ABS	English	English	TEXT	AUTO	FB	0.1593	0.2000	1797	3416
Eng-kl-dir-fb3	KL-DIR	English	English	TEXT	AUTO	FB	0.1571	0.1867	1823	3416
Eng-kl-jm-fb3	KL-JM	English	English	TEXT	AUTO	FB	0.1566	0.1917	1860	3416
Eng-tf-idf-fb2	TF-IDF	English	English	TEXT	AUTO	FB	0.1560	0.2117	1842	3416
Eng-okapi-fb3	OKAPI	English	English	TEXT	AUTO	FB	0.1540	0.1950	1733	3416
Eng-tf-idf-fb1	TF-IDF	English	English	TEXT	AUTO	FB	0.1540	0.2133	1750	3416
Eng-okapi-fb1	OKAPI	English	English	TEXT	AUTO	FB	0.1492	0.2000	1726	3416
Eng-kl-abs	KL-ABS	English	English	TEXT	AUTO	NOFB	0.1455	0.1883	1570	3416
Eng-okapi	OKAPI	English	English	TEXT	AUTO	NOFB	0.1437	0.1850	1556	3416
Eng-kl-jm	KL-JM	English	English	TEXT	AUTO	NOFB	0.1428	0.1850	1547	3416
Eng-kl-dir	KL-DIR	English	English	TEXT	AUTO	NOFB	0.1419	0.1850	1554	3416
Eng-tf-idf	TF-IDF	English	English	TEXT	AUTO	NOFB	0.1341	0.1900	1539	3416

“TF-IDF” and “OKAPI” are two typical retrieval methods, “KL” denotes Kullback-Leibler divergence based model, “DIR” denotes the smoothing technique using the Dirichlet priors, “ABS” denotes the smoothing using the absolute discounting, and “JM” denotes the Jelinek-Mercer smoothing approach.

also found that the retrieval methods by KL-divergence language models do not always outperform the traditional TF-IDF and Okapi approaches, while the language models tend to outperform the TF-IDF and Okapi approaches on average. Further, we found that the retrieval methods with pseudo-relevance feedback (FB) consistently outperform the ones without any feedback. For example, the “Eng-kl-dir” approach is the KL-divergence language model approach using the Dirichlet priors smoothing technique without feedback, which achieved only a MAP score of 0.1419. However, by using relevance feedback, the MAP performance will be importantly improved, such as the “Eng-kl-dir-fb2” solution, which achieved a MAP score of 0.1660. Moreover, comparing several different smoothing techniques, there is no clear evidence that which smoothing technique significantly outperform the others, though the Dirichlet priors smoothing approach achieved the best MAP performance among all runs. Finally, by examining the results of previous years [8], we found that the search tasks in this year seem to be more challenging for the text-based solutions.

3.3 Cross-Language Image Retrieval

In this part, we study the bilingual image retrieval using Chinese queries and English sources. To this purpose, the first step is to translate the Chinese queries into English. In our experiment, we simply test an online translation tool offered by Google ².

Given the translated results, we conducted the experimental evaluations to examine the retrieval performance. Table 2 shows the experimental results of

² http://www.google.com/language_tools

Table 2. Evaluation for cross-language image retrieval tasks between Chinese (simplified) queries and English sources (#REL=3416)

Run ID	Method	Query	Source	Modality	RunType	QE/RF	MAP	P10	REL_RET
Chn-tf-idf-fb3	TF-IDF	Chinese	S English	TEXT	AUTO	FB	0.1574	0.2000	1874
Chn-kl-dir-fb3	KL-DIR	Chinese	S English	TEXT	AUTO	FB	0.1429	0.1650	1709
Chn-tf-idf-fb2	TF-IDF	Chinese	S English	TEXT	AUTO	FB	0.1413	0.1783	1790
Chn-kl-abs-fb3	KL-ABS	Chinese	S English	TEXT	AUTO	FB	0.1406	0.1667	1713
Chn-kl-abs-fb2	KL-ABS	Chinese	S English	TEXT	AUTO	FB	0.1385	0.1500	1732
Chn-kl-dir-fb2	KL-DIR	Chinese	S English	TEXT	AUTO	FB	0.1382	0.1600	1763
Chn-kl-jm-fb2	KL-JM	Chinese	S English	TEXT	AUTO	FB	0.1380	0.1533	1801
Chn-kl-jm-fb3	KL-JM	Chinese	S English	TEXT	AUTO	FB	0.1378	0.1600	1748
Chn-kl-jm-fb1	KL-JM	Chinese	S English	TEXT	AUTO	FB	0.1345	0.1533	1696
Chn-kl-dir-fb1	KL-DIR	Chinese	S English	TEXT	AUTO	FB	0.1333	0.1650	1672
Chn-okapi-fb3	OKAPI	Chinese	S English	TEXT	AUTO	FB	0.1312	0.1517	1646
Chn-kl-abs-fb1	KL-ABS	Chinese	S English	TEXT	AUTO	FB	0.1309	0.1417	1675
Chn-tf-idf-fb1	TF-IDF	Chinese	S English	TEXT	AUTO	FB	0.1286	0.1767	1553
Chn-okapi	OKAPI	Chinese	S English	TEXT	AUTO	NOFB	0.1268	0.1417	1404
Chn-kl-dir	KL-DIR	Chinese	S English	TEXT	AUTO	NOFB	0.1265	0.1467	1410
Chn-kl-abs	KL-ABS	Chinese	S English	TEXT	AUTO	NOFB	0.1264	0.1483	1411
Chn-kl-jm	KL-JM	Chinese	S English	TEXT	AUTO	NOFB	0.1252	0.1450	1415
Chn-okapi-fb1	OKAPI	Chinese	S English	TEXT	AUTO	FB	0.1237	0.1350	1654
Chn-tf-idf	TF-IDF	Chinese	S English	TEXT	AUTO	NOFB	0.1223	0.1567	1388
Chn-okapi-fb2	OKAPI	Chinese	S English	TEXT	AUTO	FB	0.1177	0.1383	1540

cross-language retrieval evaluation. From the experimental results, we found that the average retrieval performance of the bilingual retrieval tasks is less than the results of the single language image retrieval as shown in Table 1. For example, for a same retrieval method by the KL-divergence language model with the Dirichlet priors smoothing technique, the scheme “Chn-kl-dir-fb3” achieved only the MAP of 0.1429 in the bilingual retrieval task, while the same approach “Eng-kl-dir-fd3” can achieve the MAP of 0.1571 in the single language retrieval tasks. Nonetheless, the overall performance of the bilingual approach is quite impressive. In the future work, we will study other translation techniques to improve the results [9].

3.4 Cross-Media Image Retrieval

In this task we study the combination of text and visual information for cross-media image retrieval. We consider a simple combination scheme to combine the information from both the textual and visual modalities. Specifically, for a given query, we first rank the images using the language modeling techniques. We then measure the similarity of the top ranked images with respect to the sample images of the query. Finally, we combine the similarity values from both textual and visual modalities and re-rank the retrieval results based on the overall similarity scores.

In our experiment, three types of low-level visual features are engaged: color, shape, and texture [10,11]. For color features, we use the grid color moment. Each image is partitioned into 3×3 grids and three types of color moments are extracted for representing color content of each grid. Thus, an 81-dimensional color moment is adopted for the color feature. For shape features, we employ the edge direction histogram. A Canny edge detector is used to acquire the edge images and then the edge direction histogram is computed from the edges. Each histogram is quantized into 36 bins of 10 degrees each. An additional bin

is used to count the number of pixels without edge information. Hence, a 37-dimensional edge direction histogram is used for the shape feature. For texture features, we adopt the Gabor feature [12]. Each image is scaled to 64×64 . Gabor wavelet transformation is applied on the scaled image with 5 scale levels and 8 orientations, which results in 40 subimages. For each subimage, three moments are computed: mean, variance, and skewness. Thus, a 120-dimensional feature vector is adopted for the texture feature. In total, a 238-dimensional feature vector is employed to represent each of images in the testbed.

Table 3³ shows the cross-media retrieval results, in which we evaluate the influence of fusion coefficient. Specifically, the runs with ID from “Eng-kl-dir-fb2-tv1” to “Eng-kl-dir-fb2-tv9” represent the cross-media solution with the fusion coefficient from 0.1 to 0.9, respectively. The fusion coefficient here is the weight for the visual modality. From the experimental results, we can draw several observations. Firstly, we can see that the cross-media solutions improve the retrieval performance of the text-based approach for most cases with different fusion coefficients. Secondly, we found that the best MAP performance tends to be obtained when setting the fusion coefficient to 0.4. Moreover, we found that when the fusion coefficient increases, the precision of TOP 10 returned results tends to increase. This shows that when the visual modality accounts for more, the retrieval results become more accurate and relevant. This result again verifies the effectiveness of the proposed cross-media solutions.

Table 3. Evaluation for cross-media image retrieval tasks with queries of both textual and visual information (#REL=3416)

Run ID	Query	Method	Source	Modality	RunType	QE/RF	MAP	P10	REL_RET
Visual	Euclidean	Visual	Visual	VISUAL	AUTO	NO	0.0511	0.2067	883
Eng-kl-dir-fb2-tv1	KL-DIR	English	English	MIXED	AUTO	FB	0.1748	0.2317	2036
Eng-kl-dir-fb2-tv2	KL-DIR	English	English	MIXED	AUTO	FB	0.1789	0.2350	2018
Eng-kl-dir-fb2-tv3	KL-DIR	English	English	MIXED	AUTO	FB	0.1805	0.2400	1990
Eng-kl-dir-fb2-tv4	KL-DIR	English	English	MIXED	AUTO	FB	0.1811	0.2567	1954
Eng-kl-dir-fb2-tv5	KL-DIR	English	English	MIXED	AUTO	FB	0.1794	0.2583	1900
Eng-kl-dir-fb2-tv6	KL-DIR	English	English	MIXED	AUTO	FB	0.1776	0.2883	1807
Eng-kl-dir-fb2-tv7	KL-DIR	English	English	MIXED	AUTO	FB	0.1709	0.3183	1691
Eng-kl-dir-fb2-tv8	KL-DIR	English	English	MIXED	AUTO	FB	0.1483	0.3350	1534
Eng-kl-dir-fb2-tv9	KL-DIR	English	English	MIXED	AUTO	FB	0.0902	0.3000	1223

In future work, we will study more advanced combination methods. For example, we can train SVM classifiers with labeled images and then apply the classifiers to re-rank the top images from text retrieval. We can also study semi-supervised learning to exploit the unlabeled data for the retrieval task [13].

4 Conclusions

In this paper we reported our empirical study at the ImaegCLEF 2007 photo track. We have conducted three parts of empirical evaluations for three different purposes. One is to evaluate the language models and smoothing techniques

³ This table has been updated by fixing some bug after the official evaluation.

with applications to text image retrieval. We found that the language models approaches did not achieve significantly promising results compared as we did in the ImageCLEF2005 campaign. The main reason is that the testbed in this year is totally different from 2005. In this year, images are only associated with very short text captions, which makes the text retrieval models less effective. The second evaluation is the cross-media image retrieval by combining both textual and visual information. Promising improvements were observed in our experiments. Finally, we also examined a commercial language translation tool for the cross-language retrieval tasks and found good retrieval results. In future work, we will study more effective techniques to improve current approaches.

Acknowledgements

The work in this paper was fully supported by a university startup grant from School of Computer Engineering, Nanyang Technological University, Singapore.

References

1. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: ACM SIGIR Conference, pp. 334–342 (2001)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
3. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at trec-3. In: The Third Text REtrieval Conference (TREC-3), NIST
4. Shannon, C.E.: Prediction and entropy of printed English. *Bell Sys. Tech. Jour.* 30, 51–64 (1951)
5. Zhai, C., Lafferty, J.: Model-based feedback in the kl-divergence retrieval model. In: Proc. of Tenth International Conference on Information and Knowledge Management (CIKM 2001), pp. 403–410 (2001)
6. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Chichester (1991)
7. Jelinek, F., Mercer, R.: Interpolated estimation of markov source parameters from sparse data. *Pattern Recognition in Practice*, 381–402 (1980)
8. Hoi, S.C.H., Zhu, J., Lyu, M.R.: CUHK at imageclef 2005: Cross-language and cross-media 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. 602–611. Springer, Heidelberg (2006)
9. Trujillo, A.: Translation Engines: Techniques for Machine Translation. Springer, London (1999)
10. Hoi, S.C.H., Lyu, M.R., Jin, R.: A unified log-based relevance feedback scheme for image retrieval. *IEEE Trans. on Knowledge and Data Engineering* 18(4), 509–524 (2006)
11. Wong, Y.M., Hoi, S.C.H., Lyu, M.R.: An empirical study on large-scale content-based image retrieval. In: Proc. IEEE Int. Conference on Multimedia & Expo. (ICME 2007) (2007)
12. Wu, P., Manjunath, B., Newsam, S., Shin, H.: A texture descriptor for browsing and similarity retrieval. *Journal of Signal Processing: Image Communication* 16(1–2), 33–43 (2000)
13. Hoi, S.C.H., Lyu, M.R.: A semi-supervised active learning framework for image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), New York, US, June 20–25 (2005)

Towards Annotation-Based Query and Document Expansion for Image Retrieval

Hugo Jair Escalante, Carlos Hernández, Aurelio López, Heidy Marín,
Manuel Montes, Eduardo Morales, Enrique Sucar, and Luis Villaseñor

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro No. 1, 72840, Puebla, México
hugojaire@ccc.inaoep.mx

Abstract. In this paper we report results of experiments conducted with strategies for improving text-based image retrieval. The adopted strategies were evaluated in the photographic retrieval task at *ImageCLEF2007*. We propose a Web-based method for expanding textual queries with related terms. This technique was the top-ranked query expansion method among those proposed by other *ImageCLEF2007* participants. We also consider two methods for combining visual and textual information in the retrieval process: *late-fusion* and *intermedia-feedback*. The best results were obtained by combining *intermedia-feedback* and our expansion technique. The main contribution of this paper, however, is the proposal of "*annotation-based expansion*"; a novel approach that consists of using labels assigned to images (with image annotation methods) for expanding textual queries and documents. We introduce this idea and report results of initial experiments towards enhancing text-based image retrieval via image annotation. Preliminary results show that this expansion strategy could be useful for image retrieval in the near future.

1 Introduction

Text-based image retrieval (*TBIR*) consists of using textual image annotations for obtaining images from a given annotated collection; the retrieved images should be relevant to certain user information needs (queries). Under this approach image annotations and queries are considered as small text-documents that are to be compared. Commonly, a measure based on *word matching* is used for determining similarity between query and annotations [1]. The documents that are more similar to the query are returned by the *TBIR* model. This is the predominant approach for image retrieval and most Web image search engines are based on this scheme.

TBIR methods can retrieve images related to high level concepts, (places, events, people and dates), taking advantage of the textual description of the image. This approach, however, is limited because usually textual annotations are very short, complicating the retrieval task. Additionally, *TBIR* methods rely on the quality of annotations, which in most of the cases are not complete. Furthermore, *TBIR* methods do not take into account information extracted

from images, wasting useful information that could be useful for improving their accuracy.

This paper describes the participation of *INAOE-TIA*¹ in the photographic retrieval task at *ImageCLEF2007*. Our goal was to explore different methods that could help to improve accuracy of a baseline *TBIR* model. In this respect, we proposed an effective, yet simple, Web-based technique for expanding textual queries. Furthermore, we performed experiments with two widely used methods for combining visual and textual information. The main contribution of this paper, however, is the introduction of *annotation-based expansion (ABE)*; a novel approach based on image annotation for expanding textual queries and documents. Experimental results show that this strategy could be useful for image retrieval in the near future, though some issues should be addressed first.

The rest of this paper is organized as follows. In the next Section we describe the techniques we considered for improving accuracy of the *TBIR* baseline. In Section 3 we introduce the *ABE* approach. Then, in Section 4 we present experimental results of the considered methods. Finally, in Section 5 we present some conclusions and discuss future work directions.

2 Improving TBIR Performance

In order to evaluate the gain we can have by using the different proposed techniques, we implemented a baseline *TBIR* model based on the *TMG* Matlab^R toolbox [3]. After removing meta-data and useless information, the text of the captions in the *IAPR-TC12* collection was indexed separately for the four target languages² (English, Spanish, German and Random). For indexing we used a *tf-idf* weighting, English stop words were removed and standard stemming was applied [1,3]. Queries for the baseline runs were created by using the text in topics as provided by the organizers of *ImageCLEF2007* [2] (after removing meta-data). For multilingual experiments queries were translated using the online Systran³ translation software. For retrieval we considered the cosine similarity function [1]. In the rest of this Section we present three strategies for improving accuracy of our baseline *TBIR* model.

2.1 Web Based Query Expansion

The Web is the largest repository of information that ever existed; comprising millions of documents, the Web is a very useful source of knowledge. For this reason we consider it in this work by proposing a web-based query expansion technique. The goal is to obtain (and to incorporate) related-context terms, extracted from the Web, according to the original query. The intuitive idea is

¹ Research group on machine learning, image processing and information retrieval at INAOE (<http://ccc.inaoep.mx/~tia>)

² For further details about the collection, query-target languages and the photographic retrieval task we refer the reader to the respective overview paper [2].

³ <http://www.systranbox.com/>

that expanded queries could be helpful for reaching relevant documents that may contain terms other than the ones contained in the original queries.

For each topic, we take the textual description and submitted a web-search using the Google^R search engine; the top- k snippets returned by the search engine are considered for expanding a query. We tried two approaches that we called *naïve* and *repetition*. The *naïve* approach (*NQE*) consists of taking the snippets as they are returned by Google^R with no preprocessing. On the other hand, the *repetition* approach (*RQE*) consists of retaining the most frequent terms in the set of k -snippets.

2.2 Intermedia Relevance Feedback

Intermedia feedback⁴ (*IMFB*) is a novel technique based on blind relevance feedback that has been proposed for image retrieval from annotated collections [4]. This technique consists of using a *content-based image retrieval*⁵ (*CBIR*) model with a query image for retrieving documents. The top- n documents returned are assumed to be relevant and the captions of such documents are combined to create a textual query. The textual query is then used with a *TBIR* model, and the documents returned by such a model are returned to the user. Note that the final textual query can be generated by considering different strategies. In this work we just concatenated the captions of the pseudo-relevant images. There are several variants of the method [4], some of which are published in this proceedings (see Chang et al and Clinchant et al). We tried combined runs of query expansion and *IMFB*, in which we applied first the query expansion technique and then the expanded queries were combined with the captions of the top- n relevant documents, according to the *CBIR* model, for creating the final query for the *TBIR* model. *FIRE* was used as *CBIR* system; using the baseline run provided by the *ImageCLEF2007* organizers [2].

2.3 Late Fusion of Independent Systems

Another way of enhancing *TBIR* accuracy is by adopting another well known mixed retrieval method, late fusion of independent retrievers (*LF*). This method consists of running two retrieval systems using a single (different) modality each. Then, the relevant documents returned by both systems are combined. For this work we adopted a fusion strategy based on the rank of documents according to two different systems we considered. Let T_R being the list of relevant documents, to a textual query, according to our *TBIR* model; documents are ranked in descending order of their relevance. Similarly, let V_R being the list of ranked relevant documents according to a *CBIR* system that uses the topic images as queries. We combined and re-ranked the documents returned by both retrieval systems, generating a new list of relevant documents $LF_R = \{T_R \cup V_R\}$; where

⁴ Also known as *media mapping* or *transmedia re-ranking*.

⁵ In a *CBIR* model, retrieval is done by considering images only. Note that the *IMFB* can start from text, obtaining query images for a *CBIR* system, as well.

each document $d_i \in LF_R$ is ranked according to the score formula given by Equation (1)

$$score(d_i) = \frac{\alpha \times R_{T_R}(d_i) + (1 - \alpha) \times R_{V_R}(d_i)}{1_{T_R}(d_i) + 1_{V_R}(d_i)} \quad (1)$$

where $R_{T_R}(d_i)$ and $R_{V_R}(d_i)$ is the position in the ranked list of document d_i according to the *TBIR* and *CBIR* models, respectively. $1_{T_R}(d_i)$ and $1_{V_R}(d_i)$ are indicator functions that take the value 1 if document d_i is in the list of relevant documents according to the *TBIR* and *CBIR* models respectively, and zero otherwise. The denominator accounts for documents appearing in both lists of relevant documents (T_R and V_R). Documents are sorted in ascending order of their score. Intuitively with this score documents appearing in both sets (visual and textual) will appear at the top of the ranking, considering their position in the independent lists of relevant documents. We tried several values for α and the best results were obtained with $\alpha = 0.9$.

3 Annotation-Based Document and Query Expansion

The task of automatic image annotation (*AIA*) consists of assigning textual descriptors (labels) to images (or segments in images), starting from visual attributes extracted from them. *AIA* methods are well suited for un-annotated image collections, where no textual description of the images is available. Usually, after annotation, the generated labels are used for *TBIR*. In this work, however, we propose using *AIA* methods in an already annotated collection, with the goal of expanding the textual queries and/or initial annotations with labels obtained from the content of images. While manual annotations provide semantic information that may not be obtained from the visual content of the image (when/where the picture was taken?, who took the photo?, etcetera); labels obtained with *AIA* (that is, automatic annotations) can provide information about the visual content of the image that may not be explicit in the annotation (are there *sky*, *trees*, *clouds* or *water* in the image?). In consequence both type of annotations are complementary, and this is the basis for *ABE*.

We decided to use region-level *AIA* methods for obtaining the automatic annotations. Region-level methods can provide accurate annotations and spatial context can be used for improving annotation accuracy [6]. The process we followed for *ABE* includes: (i) segmentation and feature extraction, (ii) creating a training set of annotated regions, (iii) building a classifier, (iv) testing it and expanding queries and/or documents. For segmenting the *IAPR-TC 12 benchmark* collection we used the normalized cuts algorithm [5]; which has been used by most of the region-level annotation approaches. In Figure 1 sample images segmented with normalized cuts are shown. As we can see the algorithm works well for some images, isolating single objects; however, for other images, segmentation is not as good, partitioning single objects into several regions.

After segmentation, visual attributes were extracted from each region. Attributes include color, texture and shape information of the regions (30 attributes). Each region is described by its vector of attributes. Hereafter we refer



Fig. 1. Sample images from the *IAPR TC-12* collection, segmented with normalized cuts. Manual annotations are shown for each region.

to the *attributes vector describing a region* simply by the term *region*. After feature extraction we manually annotated a set of around 2% of the total number of regions. The set of labels that can be assigned to regions was defined subjectively by the authors, by looking at the *ImageCLEF2007* textual topic descriptions. The vocabulary of labels is the following (together with the number of regions, in our training set, annotated with each label): *sky* (344), *person* (285), *building* (180), *trees* (175), *clouds* (170), *grass* (138), *water* (135), *mountain* (122), *sand* (98), *other* (55), *furniture* (47), *road* (41), *animal* (28), *snow*(25), *rock* (17), *sun* (16), *vehicle* (16), *boat* (14), *church* (9), *tower* (8), *plate* (7), *flag* (4), *statue* (4), *swimming-pool* (0). Some labels represent several concepts, for example, the label *water* was used for labeling regions of rivers, ocean, and sea. While other labels represent specific objects, such as *swimming-pool* and *tower*. We can see that there are several labels that have many training examples (for example, *Sky*, *Person*), though several other labels have only a few. This fact together with poor segmentation complicated the process of annotation.

The training set of region-label pairs is used with a *knn* classifier for annotating the un-annotated regions from the rest of the images. Note that the training set size is very small for achieving good results with the *knn* algorithm. In order to overcome, in part, the issues of poor segmentation and an imbalanced and small training set, we decided to apply a postprocessing to *knn* for improving annotation accuracy. Recently a method, called Markov random field improver (*MRFI*), for improving accuracy on *AIA* has been proposed [6]. *MRFI* considers a set of candidate labels for each region and selects an unique label for each region based on a Markov random field model that considers spatial information, labels association and the confidence of the *AIA* method on each label. We applied *MRFI* as postprocessing to *knn*.

For document expansion we annotated the 20,000 images, and expanded the original annotation with the automatic one. For query expansion we annotated the topic images and expanded the textual topics with the automatic annotations. In Figure 2 an expanded topic is shown (left) , as well as an expanded document (right). As we can see, some labels are repeated on the expanded topic (*sky*, *people* and *tree*); we considered repeated labels in order to have an impact in the *tf-idf* weighting, (that is, repeated terms are considered more representative of the query).



Fig. 2. Left: expansion of the topic 36 using annotations. Right: A sample document expanded with *ABE*. Automatic annotations are shown below each segmented image. The expanded query/document is shown below images annotations.

4 Experimental Results

A total of 95 runs were submitted to *ImageCLEF2007* comprising all of the target languages and most of the query ones. The above described methods were tested, some runs are a combination of these methods. Our top ranked entries for each language configuration together with a brief description of the methods used are shown in Table 1.

Table 1. Top ranked entries for each of the query-target language configurations comprised in the *TIA*'s submitted runs. In marked runs (*) *TIA* was the only participant group. The last column shows the percentage of improvement over the respective (monolingual) baseline *TBIR* model.

Run-ID	Languages	Methods	Type	MAP	Ranking	Improvement (%)
1	English-English	NQE+IMFB	Mixed	0.1986	22 / 142	43.3
2	Dutch-English*	NQE+IMFB	Mixed	0.1986	1 / 4	43.3
3	French-English	NQE+IMFB	Mixed	0.1986	3 / 21	43.3
4	German-English	NQE+IMFB	Mixed	0.1986	3 / 20	43.3
5	Italian-English	NQE+IMFB	Mixed	0.1986	3 / 10	43.3
6	Japanese-English	NQE+IMFB	Mixed	0.1986	2 / 6	43.3
7	Portuguese-English	NQE+IMFB	Mixed	0.1986	2 / 9	43.3
8	Russian-English	NQE+IMFB	Mixed	0.1986	2 / 6	43.3
9	Spanish-English	NQE+IMFB	Mixed	0.1986	2 / 9	43.3
10	Visual-English*	NQE+ABE+IMFB	Mixed	0.1925	1 / 1	38.9
11	German-German	NQE+LF	Mixed	0.1341	13 / 30	44.5
12	English-German	NQE+LF	Mixed	0.1113	11 / 17	19.9
13	Spanish-Spanish	NQE+LF	Mixed	0.1481	5 / 15	7.71
14	English-Spanish	NQE+LF	Mixed	0.1145	2 / 6	-16.7
15	Dutch-Random*	NQE	Text	0.0828	1 / 2	10.2
16	English-Random	NQE+IMFB	Mixed	0.1243	6 / 11	65.5
17	French-Random	NQE+IMFB	Mixed	0.1243	3 / 10	65.5
18	German-Random	NQE+IMFB	Mixed	0.1243	4 / 11	65.5
19	Italian-Random*	NQE	Text	0.0798	1 / 2	6.26
20	Portuguese-Random*	NQE	Text	0.0296	1 / 2	-60.4
21	Russian-Random*	NQE	Text	0.0763	1 / 2	1.6
22	Spanish-Random*	NQE+ IMFB	Mixed	0.1243	1 / 5	65.5

As we can see, most of the entries are ranked near the first one, and most of them outperform significantly the *TBIR* baseline (column 7). The larger improvement is of around 65%, which is a significant improvement over the *TBIR*

baseline. We had some negative results, though we should emphasize that all runs (including bilingual) are compared to a monolingual *TBIR* model. For example the 14th run was compared to a Spanish-Spanish *TBIR* model. It is clear that translation mistakes can degrade the performance in these runs.

The best performance overall runs was obtained by using *IMFB* together with *NQEs*. Actually the *NQE* is present in all of the top ranked runs. *NQE* outperformed *RQE* in all of the language configurations, and according to the official results *NQE* was the best technique among those other proposed for query expansion. This is a surprising result because with *NQE* several noisy terms are added to the queries. While with *RQE* only the terms that most appear among all the snippets are added. The good results of *NQE* are due to the inclusion of many highly related terms, while the insertion of some noisy terms does not affect the performance of the retrieval model.

We can observe that the runs with *IMFB+NQE* for target language English have exactly the same *MAP* value, independently of the query language. This means that the generated queries were dominated by *IMFB*. *IMFB* outperformed the *LF* method in all of the runs if we consider the *MAP*. However, an interesting finding is that *LF* obtained higher recall than any other method we tried, retrieving 16% more documents than *IMFB*. This means that the ranking strategy we adopted for *LF* should be improved.

Six runs based on *ABE* were submitted to the *ImageCLEF2007*. In these runs document and query expansion were combined with the other techniques proposed in previous sections. The descriptions of the annotation based expansion (*ABE*) runs submitted to *ImageCLEF2007* are shown in Table 2. Run 1 in Table 2 is the same as run 10 in Table 1. This is an interesting run because we start from query images only, and by *ABE* and *IMFB* we build a textual query that is used with a *TBIR* model. This approach is language independent as it starts from images only, therefore, it could be very helpful for cross-lingual image retrieval. This was the only run for the language configuration visual-English.

Results with *ABE* are mixed. The two top ranked runs with *ABE* correspond to entries that used *ABE+IMFB*. One should note that with *ABE* we have an insignificant loss of accuracy. In consequence, the favorable result is due to the *IMFB* performance instead of the *ABE* technique. The third *ABE* ranked run used *ABE* of documents and queries with *NQE+LF* which obtained a slightly

Table 2. Settings of the *ABE* runs. An **X** indicates that the corresponding technique is used. *ABQE* is for *annotation-based query expansion* and *ABDE* is for *annotation-based document expansion*. The ranking position is shown. *Diff.* is the accuracy we gain-loss with respect of using only the *methods* of column 2 without *ABE*. The last column show the percentage of improvement with respect to the *TBIR* baseline.

ID	Methods	ABQE	ABDE	Rank	Diff.	Imp. (%)
1	Baseline,IMFB	X	-	57	-0.006	38.9
2	Baseline,IMFB	X	X	58	-0.006	38.9
3	NQE,LF,	X	X	84	-0.001	22
4	NQE,Baseline	X	X	133	-0.001	11.7
5	NQE,LF	X	-	389	-0.092	-44.1
6	Baseline	X	X	447	-0.111	-79.5

lower *MAP* than *NQE+LF* without *ABE*. Therefore no gain can be attributed to the *ABE* technique. The other *ABE* runs were ranked low. We should emphasize that this was our very first effort towards developing annotation based methods for improving image retrieval. Several issues should be addressed first in order to evaluate the added value of *ABE*, these are: using better segmentation tools, creating a large and balanced training set of annotated regions, defining a better suited vocabulary for annotation and trying other *AIA* methods instead of *knn*.

5 Conclusions

We have presented experimental results obtained with different strategies for improving *TBIR* methods. An effective Web expansion method was proposed and we tried two widely known mixed retrieval methods. Furthermore, we proposed *ABE* and performed initial experiments with this technique. Experimental results give evidence that most of the methods we considered improved accuracy of a *TBIR* baseline (up to 65%). The best runs were those based on *IMFB+NQE*. The *NQE* method was the top ranked query expansion method among those proposed by other participants. *IMFB* outperformed *LF* by a large margin in *MAP*, though *LF* obtained higher recall. Results with *ABE* give evidence that *AIA* methods could be helpful for image retrieval from annotated collections. This because promising results were obtained even when segmentation was poor, the training set was extremely small and imbalanced, annotations did not covered the objects present within the image collection and we used a very simple classifier. For future work we will address all of these issues and we will perform extensive experimentation for evaluating the advantages/disadvantages of *ABE*.

Acknowledgements. We would like to thank all of the organizers of *Image-CLEF2007*. This work was partially supported by CONACyT under project grant 61335.

References

1. Baeza, R., Ribeiro, B.: Modern Information Retrieval. Pearson E. L (1999)
2. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the CLEF (2007)
3. Zeimpekis, D., Gallopoulos, E.: TMG: A MATLAB Toolbox for generating term-document matrices from text collections. Grouping Multidimensional Data: Recent Advances in Clustering, 187–210 (2005)
4. Chang, Y., Chen, H.: Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval. In: Working Notes of the CLEF (2006)
5. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Patt. An. and Mach. Intell. 22, 888–905 (2000)
6. Escalante, H., Montes, M., Sucar, E.: Word Co-occurrence and MRF's for Improving Automatic Image Annotation. In: Proc. of the 18th BMVC (2007)

Content-Based Image Retrieval Using Combined 2D Attribute Pattern Spectra

Florence Tushabe and Michael. H.F. Wilkinson

Institute of Mathematics and Computing Science, University of Groningen,
P. O. Box 407, 9700 AK Groningen, The Netherlands
`m.h.f.wilkinson@rug.nl`, `florence@cs.rug.nl`

Abstract. This work proposes a region-based shape signature that uses a combination of three different types of pattern spectra. The proposed method is inspired by the connected shape filter proposed by Urbach et al. We extract pattern spectra from the red, green and blue color bands of an image then incorporate machine learning techniques for application in photographic image retrieval. Our experiments show that the combined pattern spectrum gives an improvement of approximately 30% in terms of mean average precision and precision at 20 with respect to Urbach et al's method.

1 Introduction

The most popular content-based image retrieval descriptors follow the standard MPEG-7 visual tool-set [1]. They include descriptors based on color, texture, shape, motion and localisation. We test an alternative method of obtaining the image descriptor by the application of granulometric operations and machine learning techniques. Granulometric operations are applied to the image at different scales and levels of complexity to derive information about the distribution of its contents[2]. Attribute filtering is a relatively new and efficient way of implementing granulometry. Desired descriptors like size, spatial location or shape can be well represented with appropriate attributes like area [3], moments [4,5] or shape [6]. A size granulometry for example uses sieves of increasing sizes to obtain the size distribution of the image. Previous works like [7,8] use a structuring element approach for the granulometric operations. However, recent studies have found connected filtering to be faster and equal or sometimes better in performance than the SE approach [6]. In [6], a shape filter of a 2-D pattern spectrum consisting of an area and non-compactness spectrum is proposed. We extend the shape spectrum proposed in [6] and apply it to a photographic data set containing everyday vacation pictures [9]. This is because most shape-based image retrieval studies concentrate on artificial images or highly specialised domain-specific image data sets. The proposed shape spectrum consists of three rotation and scale invariant spectra: the area-non-compactness [10], area-compactness and area-entropy pattern spectra. They are weighted, combined and used for image retrieval within large-scale databases. The rest of the paper is organised

as follows: Section 2 briefly describes the theory of the method employed, Section 3 contains the experimental set-up and Sections 4 and 5 give the results, discussions and concluding remarks.

2 Theory

Connected attribute filtering decomposes an image into sets of connected components. Each component adopts a single attribute value, r , and is considered for further processing only when r satisfies a given criterion. Attribute filtering is manifested through attribute openings or thinnings and is extensively discussed in [11]. Let C, D be connected components of set X and Ψ a binary image operator. Attribute openings are characterized by being increasing ($C \subseteq D \Rightarrow \Psi(C) \subseteq \Psi(D)$), idempotent ($\Psi\Psi(C) = \Psi(C)$) and anti-extensive ($\Psi(C) \subseteq C$). Example attributes include area, perimeter and moment-of-inertia. On the other hand, attribute thinnings are characterised by being idempotent, anti-extensive and non-increasing ($C \subseteq D \not\Rightarrow \Psi(C) \subseteq \Psi(D)$). Example attributes are length, compactness, non-compactness, circularity and entropy.

If X, Y represent an image, then the size granulometry (Γ_r) is a set of filters $\{\Gamma_r\}$ with r from some totally ordered set Λ (usually $\Lambda \subset \mathbb{R}$ or $\mathbb{Z}$) satisfying the properties:

$$\Gamma_r(X) \subseteq X \quad (1)$$

$$X \subseteq Y \Rightarrow \Gamma_r(X) \subseteq \Gamma_r(Y) \quad (2)$$

$$\Gamma_s(\Gamma_r(X)) = \Gamma_{\max(r,s)}(X) \quad (3)$$

$$\forall r, s \in \Lambda$$

Breen and Jones [11] show that attribute openings indeed provide size granulometries since equations (1),(2) and (3) define Γ_r as being anti-extensive, increasing and idempotent respectively. Similarly, Urbach and Wilkinson [10] show that a shape granulometry can be obtained from attribute thinnings. The shape granulometry, of X , is a family of filters, $\{\Phi_r\}$, with shape parameter, r , from some totally ordered set Λ (usually $\Lambda \subset \mathbb{R}$ or $\mathbb{Z}$) with the following properties:

$$\Phi_r(X) \subseteq X \quad (4)$$

$$\Phi_r(tX) = t(\Phi_r(X)) \quad (5)$$

$$\Phi_s(\Phi_r(X)) = \Phi_{\max(r,s)}(X) \quad (6)$$

$$\forall r, s \in \Lambda \text{ and } t > 0$$

Equations (5),(6) and (7) define Φ_r as anti-extensive, scale invariant and idempotent respectively.

2.1 2-D Pattern Spectra

The results of the application of granulometry to an image can be stored in a pattern spectrum [3]. A 2D pattern spectrum represents the results of two granulometric operations in a single 2-dimensional histogram. The shape filter proposed in this work consists of a size-shape pattern spectrum.

The size pattern spectrum, $s_\Gamma(X)$, obtained by applying the size granulometry, $\{\Gamma\tau\}$, to a binary image X is defined by [3] as:

$$(s_\Gamma(X))(u) = - \left. \frac{dA(\Gamma_r(X))}{dr} \right|_{r=u} \quad (7)$$

where $A(X)$ is the area of X .

While the shape pattern spectrum, $s_\Phi(X)$, is obtained by applying the shape granulometry, $\{\Phi\tau\}$, to a binary image X and defined by [6] as:

$$(s_\Phi(X))(u) = - \left. \frac{dA(\Phi_r(X))}{dr} \right|_{r=u} \quad (8)$$

where the difference with (7) is in the use of the shape granulometry.

2.2 Computing the Pattern Spectra

The max-tree approach [12,13] was used to implement the attribute thinnings and openings. Let the peak components, P_h^k of an image represent the connected components of the threshold set at gray level h with k from some arbitrary index set. These peak components are arranged into a tree structure and filtered by removing nodes whose attribute values are less than a pre-defined threshold, T . Thus, the max tree is a rooted tree in which each of its nodes, C_h^k , at gray-level h corresponds to a peak component, P_h^k [13]. An example is shown in Figure 1 which illustrates the peak components, P_h^k , of a 1-D signal, the corresponding C_h^k at levels $h = 0, 1, 2, 3$, the resultant max-tree and corresponding spectrum. Note that two attributes are shown per node, the first of which is the size attribute which increases as the tree is descended. The second attribute, which is the shape attribute is not increasing.

The method of generating the 2D spectrum has been adopted from Urbach et al [6]. Let $\{\Gamma_r\}$ be a size distribution with r from some finite set Λ_r and $\{\Phi_s\}$ a shape distribution with s from some index set Λ_s . If S is the 2-D array that stores the final 2-D spectrum, then each cell, $S(r, s)$, contains the sum of gray levels of C_h^k that falls within size class r — and r and shape class s — and s . The 2-D pattern spectrum is then computed from the max-tree as follows:

- Set all elements of the array S to zero.
- Compute the max-tree according to the algorithm in [12].
- As the max-tree is built, compute the area $A(P_h^k)$, perimeter $P(P_h^k)$, histogram of the gray levels and moment of inertia $I(P_h^k)$ of each node.
- For each node C_h^k :

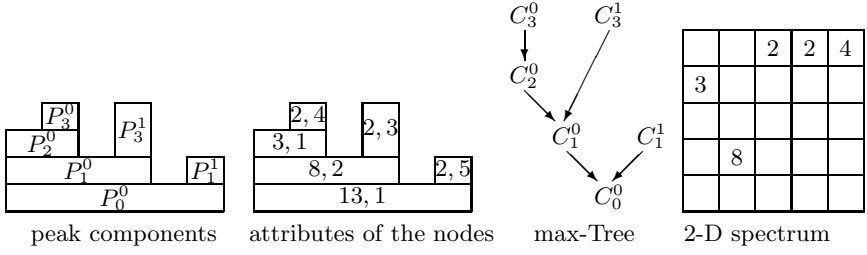


Fig. 1. Peak components(P_h^k), their attributes, corresponding (C_h^k) (the max-tree) and the resulting pattern spectrum (right)

- Compute the size class r from the area of P_h^k .
- Compute the shape class s from the shape attribute of P_h^k .
- Compute the gray level difference δ_h , between the current node and its parent;
- Add the product of δ_h and $A(P_h^k)$ to $S(r, s)$.

The shape attributes chosen are non-compactness, N , defined as

$$N = \frac{I(P_h^k)}{A^2(P_h^k)}, \quad (9)$$

compactness, C , defined as

$$C = \frac{P^2(P_h^k)}{A(P_h^k)}, \quad (10)$$

and finally, Shannon entropy

$$H = - \sum p(i) \log_2 p(i), \quad (11)$$

with $p(i)$ the probability with which gray level i occurs in P_h^k .

3 Experiments

The objective of our experiments was: given a sample image, find as many relevant images as possible from the IAPR TC-12 photographic collection [14]. Our method uses the three query images that were provided per topic. The detailed methodology is as follows:

1. Separate the jpeg image into three different images, each representing its red, green and blue color bands. This is after initial analysis shows that RGB representation improves results unlike YUV and XYZ which performed worse than not separating the images.
2. Extract the desired pattern spectra from all the images including the query images. A 20 by 15 bin histogram that eventually translates into a 1×600 array representation was chosen. When concatenated, the spectra retrieved from the three color bands forms a 1×1800 vector per spectrum type. The three spectra that were tested are:

- (a) Area and Non-Compactness (A-N) spectrum
 - Area: Is a size filter that represents the number of pixels in the component. Initial experiments in [15] showed that the discriminative power lies more in the larger particles rather than the smaller ones. Therefore all particles less than 30% of the total image size were ignored.
 - Non-Compactness: Thresholds of 1 - 53 were used for the non-compactness attribute since it gave the best MAP when compared with other thresholds ranging between $T = 1 : 100$.
 - (b) Area and Compactness (A-C) spectrum
 - Area: Same thresholds as above.
 - Compactness: The thresholds chosen for compactness are $T = 600$ since it registered the highest MAP when compared with other thresholds ranging between $T = 1 : 1000$.
 - (c) Area-Entropy (A-E) spectrum
 - Area: Same thresholds as above.
 - Entropy: A threshold of $T = 8$ was chosen because it is the maximum entropy that any component can achieve.
3. The spectra were separated into two equal parts, A and B , referring to larger and smaller features in the images.
 4. The baseline distance, $d_{x,j}$, of any two images x and j is given by:

$$d_{x,j} = w_a d_{A(x,j)} + w_b d_{B(x,j)} \quad (12)$$

where $w_{a,b}$ are the weights of parts A and B of the spectrum and $d_\alpha(x, j)$ the L1 norm distance of image x and j as computed from attribute α of the spectrum. The weights chosen for area - non-compactness is $w_a = 0.7$ and $w_b = 0.3$; area - compactness is $w_a = 0.7$ and $w_b = 0.3$; and area - entropy attributes $w_a = 0.5$ and $w_b = 0.5$. These weights were found by trial and error.

5. The 250 most significant features from the 1×1800 spectra are selected and used to train the query images using the naive bayesian classifier from [17,16]. The images are then classified by each of the spectra into classes consisting of the 60 topics. The distance, d_x , of an image from a particular topic is reduced by a given percentage, p if it has been classified within that topic. This is done because we wish to obtain a single distance, and the bayesian classifier from [17] works with a different distance measure than d_x . Parameter p is the classification weight and is 20% for for A-N and A-E and 70% for A-C feature sets respectively. These percentages were also determined empirically.
6. The distance, D_x of X from topic T is the minimum of its distances from the three topic images. The final distance of image x from topic image y is the weighted addition of its distances from the three spectra.

$$D_x = \min_{j \in T} \{0.75d_{x,j}^N + 0.20d_{x,j}^C + 0.05d_{x,j}^H\} \quad (13)$$

where $d_{x,j}^N$, $d_{x,j}^C$, $d_{x,j}^H$ are the distances between x and j depending on their A-N, A-C and A-E spectra, respectively.

7. The similarity measure between images X and Y is then calculated using:

$$Sim(X, Y) = 1 - \frac{D_x}{D_{\max}} \quad (14)$$

where $D_{\max}$ is the maximum of D_x over the data set, which helps in normalising D_x .

4 Results

The experiments were implemented in C and matlab and run on an AMD Opteron-based machine. Feature extraction took approximately 3 seconds per image. The overall performance of this method has shown that combining the three spectra improves the MAP of the best performing single spectrum by over 28%. Table 1 gives the detailed results of the different combinations that were performed. They show that the A-N spectrum has the highest discriminative

Table 1. Performance of the spectra

Run	MAP	P20	Relevant	% improvement
A-N	0.0444	0.1258	830	-
A-C	0.0338	0.1100	819	-
A-E	0.0265	0.0767	622	-
A-N and A-C	0.0539	0.1508	932	21.4
A-N and A-E	0.0479	0.1358	846	7.9
A-N, A-C and A-E	0.0571	0.1608	926	28.6

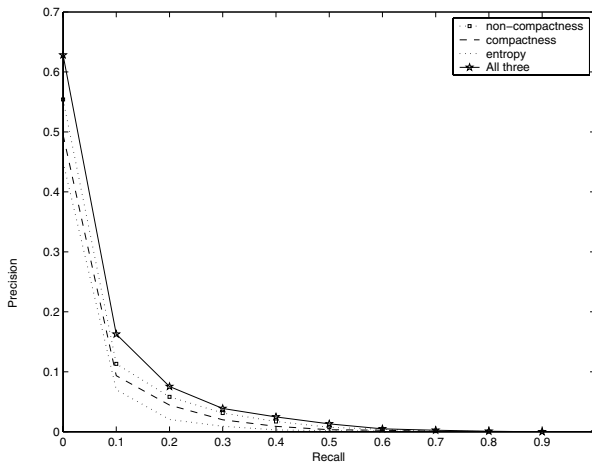


Fig. 2. Interpolated Precision - Recall Averages

power, followed by A-C and A-E respectively. Figure 2 illustrates the interpolated precision-recall average for the three separate and the combined spectra. As expected, at any given point, the precision of the combined spectrum is much higher than any of the individual ones. Initial results showed that Bayes classification out-performed k-nearest neighbour and decision tree. Bayes classification improves the MAP of the combined filter by 28% from 0.0444 to 0.0571 and precision at 20 from 0.1258 to 0.1333.

5 Discussion

Our experiments have shown that using only one technique, i.e, the 2D pattern spectra, produces very promising results for CBIR. There is no doubt that combining it with other visual descriptors like color or texture will further enhance performance for image retrieval. This work proposes a feature vector that combines three 2D pattern spectra: the area-non-compactness, area-compactness and area-entropy spectra. The combined spectrum translates in an improved performance in terms of both the mean average precision and precision at 20. Given the small training set used and simple retrieval scheme, the registered performance indicates that this feature set shows promise and should be developed further. Urbach et al [6] already showed that the area-non-compactness spectrum is very robust against noise in the application of diatom identification. The difference in performance between the different pattern spectra may be attributed to differences in robustness to noise. Compactness is probably less robust through the use of the perimeter parameter. The fact that the A-C spectrum required a classification weight of 70% compared to 20% for A-N and A-E respectively could indicate that the decision boundary with the simple nearest neighbor classifier is less reliable in the case of compactness. The relatively poor performance of entropy may mean that shape is relatively more important than variation in gray level. We believe that choosing features using more advanced relevance learning techniques [18,19] as well as using a larger training set will enhance the MAP scores registered here. Secondly, obtaining the spectra from specific objects (cartoon) as opposed to the whole image can also be tried out [20,21]. Further advancements should include relevance feedback by users.

References

1. Bober, M.: MPEG-7 Visual Descriptors. *IEEE Transactions in Circuits and Systems for Video Technology* 11(7), 703–715 (2001)
2. Matheron, G.: *Random sets and integral Geometry*. John Wiley and Sons, Chichester (1975)
3. Maragos, P.: Pattern Spectrum and Multiscale shape Representation. *IEEE Transactions Pattern Analy. Mach. Intel.* 11(7), 701–715 (1989)
4. Wilkinson, M.H.F.: Generalized Pattern Spectra sensitive to Spatial Information. In: *Proceeding of the 16th International Conference on Pattern Recognition*, Quebec City, vol. 1, pp. 701–715 (2002)

5. Hu, M.K.: Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory* IT-8, 179–187 (1962)
6. Urbach., E.R., Roerdink., J.B.T.M., Wilkinson, M.H.F.: Connected Shape-Size Pattern Spectra for Rotation and Scale-Invariant Classification of Gray-Scale Images. *Trans. Pattern Analy. Machine Intell.* 29(2), 272–285 (2007)
7. Bagdanov., A., Worring, M.: Granulometric analysis of document images. In: *Proceeding of the 16th International Conference on Pattern Recognition*, vol. 1, pp. 468–471 (2002)
8. Fuertes Garcia., J.M., Lucena Lopez, M., Gomez, J.I., de la Blanca, N.P., Fdez-Valdivia, J.: Content Based Image Retrieval Using a 2D Shape Characterization. In: *Fifth Iberoamerican Symposium On Pattern Recognition (SIARP)*, Lisbon, pp. 529–544 (2000)
9. Grubinger., M., Clough., P., Clement, L.: The IAPR TC-12 Benchmark for Visual Information Search. *IAPR Newsletter* 28(2), 10–12 (2006)
10. Urbach, E.R., Wilkinson, M.H.F.: Shape-Only Granulometries and Grey-scale shape filters. In: *International Symposium on Mathematical Morphology*, Sydney, Australia (2002)
11. Breen, E.J., Jones, R.: Attribute openings, thinnings and granulometries. *Computer Vision Image Understanding* 64(3), 377–389 (1996)
12. Salembier, P., Oliveras, A., Garrido, L.: Antiextensive connected operators for image and sequence processing. *IEEE Transactions In Image Processing* 7(4), 555–570 (1998)
13. Meijster, A., Wilkinson, M.H.F.: A comparison of Algorithms for Connected Set openings and Closings. *IEEE Trans. Pattern Analy. Mach. Intell.* 34(4), 484–494 (2002)
14. Nardi, A., Peters, C.: Working Notes of the 2007 CLEF Workshop, Budapest (2007)
15. Tushabe, F., Wilkinson, M.H.F.: Content-based Image Retrieval Using Shape-Size Pattern Spectra. In: *Working notes for the CLEF 2007 Workshop*, Hungary (2007)
16. Demsar, J., Zupan, B., Leban, G.: *Orange: From Experimental Machine Learning to Interactive Data Mining*. Faculty of Computer and Information Science, University of Ljubljana (2004)
17. Kira, K., Rendell, L.: A practical approach to feature selection. In: *Proceedings of the 9th International Conference on Machine Learning*, Aberdeen, pp. 249–256 (1992)
18. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. *Neural Networks* 15, 1059–1068 (2002)
19. Hammer, B., Strickert, M., Villmann, T.: On the generalization capability of GR-LVQ networks. *Neural Processing Letters* 21, 109–120 (2005)
20. Maragos, P., Evangelopoulos, G.: Levelling cartoons, texture energy markers and image decomposition. In: *Proceeding of the 8th International Symposium on Mathematical Morphology*, Rio de Janeiro, pp. 125–138 (2007)
21. Sofou, A., Evangelopoulos, G., Maragos, P.: Coupled Geometric and Texture PDE - based Segmentation. In: *Proceeding of the International Conference on Image Processing*, vol. II, pp. 650–653 (2005)

Text-Based Clustering of the ImageCLEFphoto Collection for Augmenting the Retrieved Results

Osama El Demerdash, Leila Kosseim, and Sabine Bergler

CLaC Laboratory - Department of Computer Science & Software Engineering -
Concordia University
{osama_el,kosseim,bergler}@cse.concordia.ca

Abstract. We present our participation in the 2007 ImageCLEF photographic ad-hoc retrieval task. Our first participation in this year's ImageCLEF comprised six runs. The main purpose of three of these runs was to evaluate the text and visual retrieval tools as well as their combination in the context of the given task. The other purpose of our participation was to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of the ImageCLEF Ad hoc task. We used the preclustered collection to augment the search results of the retrieval engines. For retrieval, we used two publicly available libraries; *Apache Lucene* for text and *LIRE* for visual retrieval. The clustered-augmented results reduced slightly the precision of the initial runs. While the aspired results have not yet been achieved, we note that the task is useful in assessing the validity of the clusters.

1 Introduction

In this paper, we present our participation in the 2007 ImageCLEF photographic ad-hoc retrieval task. The task deals with answering 60 queries of variable complexity from a repository of 20,000 photographic images in the IAPR TC-12 collection. A full description of the task and the collection can be found in [1]. Our first participation in this year's ImageCLEF comprised six runs. The main purpose of three of these runs was to evaluate the text and content-based retrieval tools in the context of the given task. We therefore would like to stress that the evaluation of these tools can only be considered under the given parameters of the task, including the queries, the image collection and our utilization of these tools.

The other purpose of our participation was to experiment with applying clustering techniques to this task, which has not been done frequently in previous editions of the ImageCLEF Ad hoc retrieval task. While this task of ImageCLEF was not intended for the evaluation of interactive methods, it could still be useful in the evaluation of certain aspects of such methods such as the validity of the initial clusters in our case.

2 Related Work

Clustering, as an unsupervised machine learning mechanism, has rarely been investigated within the context of the ImageCLEF ad-hoc Retrieval task. This

could be due to that clustering methods lend themselves more readily to interactive tasks and iterative retrieval. In the IR field, clustering has been experimented with extensively [2]. Its different applications involve clustering the whole data collection, part of it or clustering only the search results. In [3], images are clustered using labels from the surrounding HTML text. [4] applied clustering to content-based image retrieval using the Normalized Cut (NCut) algorithm under a graph representation. Another spectral clustering algorithm, *Locality Preserving Clustering (LPC)*, was introduced in [5] and found to be more efficient than NCut for image data. There is very little work in the literature on clustering using both content-based and text-based features. [6] and [7] describe successive clustering applied on text features then image features. The textual features comprised a vision based text segment as well as the link information while the *Color Texture Moments (CTM)*, a combined representation of color and texture were chosen for visual features. The only research we came across in the literature combining simultaneously image and textual features were from Microsoft Research Asia in 2005. [8] and [9] both use co-clustering techniques.

3 Resources

For retrieval, we used two publicly available libraries; *Apache Lucene* [10] for text and *LIRE* [11] for visual retrieval. Since our runs involved only English/English and Visual queries, we did not make use of any translation tools.

3.1 Text Retrieval

For text retrieval, we used the Apache Lucene engine, which implements a TF-IDF paradigm. Stop-words were removed, and the data was indexed as *field data* retaining only the *title*, *notes* and *location* fields, all of which were concatenated into one field. This helped reduce the size of the index, since our initial plan was to base the clustering on word-document cooccurrence and document-document similarity matrices. The number of indexed terms was 7577 from the 20,000 English source documents. All text query terms were joined using the *OR* operator. We did not apply any further processing of text queries.

3.2 Content-Based Retrieval

For visual retrieval, we employed v0.4 of the LIRE library which is part of the Emir/Caliph project available under the GNU GPL license. At the time of carrying out the experiments, LIRE offered three indexing options from the MPEG-7 descriptors: ScalableColor, ColorLayout and EdgeHistogram (a fourth one, Auto Color Correlogram, has since been implemented). The first two of these are color descriptors while the last is a texture one. We used all three indices. The details of these descriptors can be found in [12]. Only the best 20 images of each visual query were used. The visual queries consisted of the three images provided as example results. Thus, a maximum of 60 image results from visual queries were used in the evaluation.

4 Clustering Methodology

Three of our runs utilized preclustering of the data collection to augment the result set of the retrieval engines. Although we had intended in the beginning to cluster the results obtained from the text retrieval and content-based retrieval, we resorted to clustering the collection, given the small number of relevant results per query (compared to results from searching the World Wide Web for example).

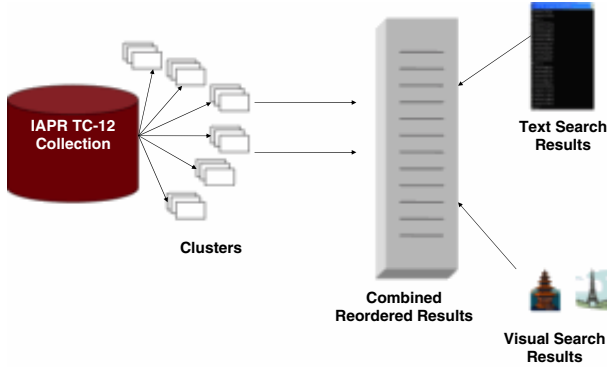


Fig. 1. Overview of the System

We employed a simple one-pass clustering algorithm which relied on forming clusters of the terms in the documents as they were processed. If a document's similarity to a cluster exceeded a certain threshold (n), this document and its new terms were added to the term/document cluster. When a document was not associated with any cluster, it was temporarily assigned its own, which was deleted in the end if no other documents were associated with it. Also clusters larger than size (s) or smaller than size (m) were discarded since they were deemed unmeaningful. We did not, however, experiment with the parameters s and n and chose them with the little intuition we had about the data. The resulting clusters overlapped and did not cover all documents.

The following parameters were used in the experiments:

- Number of top results used for cluster expansion $t = 20$
- Number of results retained from image search $r = 20$
- Minimum number of common words to be in the same cluster $n = 3$
- Minimum size of cluster $m = 3$
- Maximum size of cluster $s = 300$

Figure 1 shows an overview of the system. In the mixed run (clacTXCB), we combined the results from Lucene text search and LIRE visual search by ranking the common ones highest, followed by all other text results and finally the visual results are added at the bottom of the list. This is due to the higher confidence we had in the text search results.

For augmenting the results from the clusters, we searched the clusters for the top t results and whenever one was found we inserted the other members of the cluster at this position in the result set, taking care not to include duplicate results from different clusters.

5 Results and Analysis

We submitted six runs at ImageCLEF 2007:

- clacTX: uses Lucene for text search
- clacCB: uses LIRE for visual search
- clacTXCB: combines the results from Lucene and LIRE
- clacCLSTX: augments clacTX with clusters
- clacCLSTCB: augments clacCB with clusters
- clacCLSTXCB: augments clacTXCB with clusters.

Table 1 shows the results our runs obtained at ImageCLEFphoto 2007 as well as the average, median and best results of the track. Our highest ranked run, (clacTXCB), is the one that combined results from Lucene (text retrieval) and LIRE (visual retrieval), getting a higher MAP as well as better performance on all other measures than the other runs. For this run, we used a combined list of the results from both engines, ranking common results highest on the list as described in Section 4.

The poor performance of our text-only run (clacTX) can be mainly attributed to the absence of stemming and query expansion/feedback. The total number of terms in the text index is 7577. When using a stemmer this figure is reduced by approximately 800. The results improve by an order of 1% to 2%.

As for query expansion, we estimate that the results can improve significantly by employing geographical gazetteers as well as synonyms. Indeed, further examination of the results shows that our poorest results were obtained for queries that reflect a combination of these two factors. For example, our poorest precision was obtained for topics no. 40 (*tourist destinations in bad weather*) and 41 (*winter landscape in South America*).

Table 1. Results at ImageCLEF 2007

Run ID	Modality	MAP	P10	P20	P30	GMAP	Rel
clacTXCB	Mixed	0.1667	0.2750	0.2333	0.1599	0.0461	1763
clacCLSTXCB	Mixed	0.1520	0.2550	0.2158	0.1445	0.0397	1763
clacTX	Text	0.1355	0.2017	0.1642	0.1231	0.0109	1555
clacCLSTX	Text	0.1334	0.1900	0.1575	0.1205	0.0102	1556
clacCB	Visual	0.0298	0.1000	0.1000	0.0584	0.0058	368
clacCLSTCB	Mixed	0.0232	0.0817	0.0758	0.0445	0.0038	386
Average run	N/A	0.1292	0.2262	0.1913	0.1724	0.0354	1454
Median run	N/A	0.1327	0.2017	0.1783	0.1659	0.0302	1523
Best run	Mixed	0.3175	0.5900	0.4592	0.3839	0.1615	2251

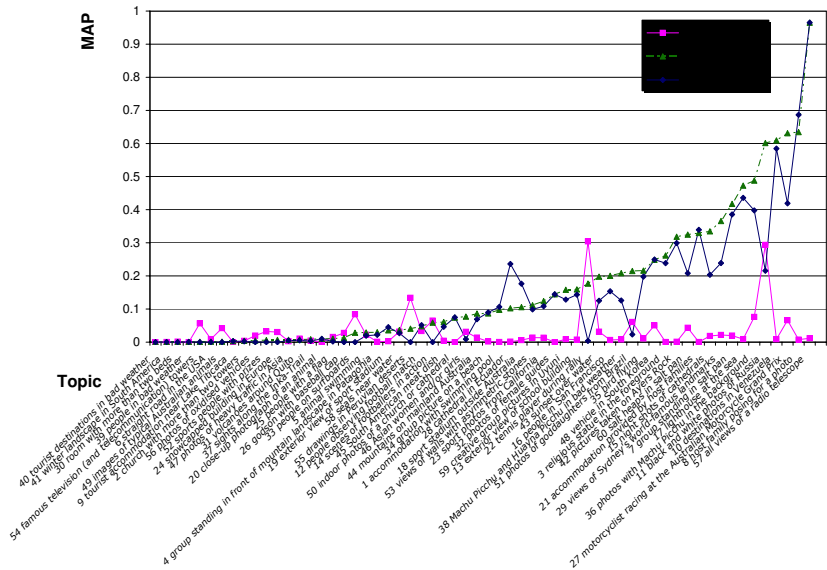


Fig. 2. Comparison between Text, Visual and Combined Results by Topic

Figure 2 shows the detailed performance by topic of the visual and text systems as well as their combination. We deduce from this figure that the run combining both systems had better results than those using the text or the visual system alone in a majority of the topics. In the few cases where the text retrieval obtained a higher MAP, the combined result was affected by noise induced from the visual results. On the other hand, the visual results achieved higher precision in some topics because of our reliance mainly on the text results, due to the higher confidence we had in them.

Our simple method of augmenting results using the preclustered data deteriorated the results in all three cases: text, visual and their combination. The main reason is that our clusters were less fine-grained than the requirements of the queries. We retained only 84 clusters of which only a handful were useful. When we experimented with the parameters we found that basing the clustering on a higher number of common words would lead to improving the results over the runs that do not employ the clusters. The one-pass clustering algorithm was unable to find this optimal parameter.

As for the other parameters described in section 4, they did not count for significant changes in the results. The number of results retained per visual query (=20) was found to be the most appropriate. Increasing or decreasing it degrades the precision. The same observation applies to the number of top documents (=20) used in augmenting the results, which can be attributed to the degrading precision after the top 20 as can be seen from the results. For the size of clusters we noted that very small clusters, which number below 30, were not useful since it is rare that one of their members happens to be in the query

results. On the other hand, large clusters (with size > 200) introduce noise and reduce the precision.

6 Conclusion and Future Work

We intend to experiment with clustering the result set as well as introducing query expansion and pseudo-relevance feedback. Our final target is clustering based on both text and visual features.

Our first participation at ImageCLEF was satisfactory in that we were able to evaluate the IR tools we chose, as well as the validity of the initial clusters produced from a simple unsupervised clustering method.

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF-photo 2007 Photographic Retrieval Task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
2. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008), Online 17/08/2007, <http://www-csli.stanford.edu/schuetze/information-retrieval-book.html>
3. Sunayama, W., Nagata, A., Yachida, M.: Image Clustering System on WWW Using Web Texts. In: Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS 2004), pp. 230–235 (2004)
4. Chen, Y., Wang, J.Z., Krovetz, R.: Content-based Image Retrieval by Clustering. In: MIR 2003: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 193–200. ACM Press, New York (2003)
5. Zheng, X., Cai, D., He, X., Ma, W.Y., Lin, X.: Locality Preserving Clustering for Image Database. In: MULTIMEDIA 2004: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 885–891. ACM Press, New York (2004)
6. Cai, D., He, X., Ma, W.Y., Wen, J.R., Zhang, H.: Organizing WWW Images Based on the Analysis of Page Layout and Web Link Structure. In: Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, Taipei, Taiwan, 27–30 June 2004, pp. 113–116. IEEE, Los Alamitos (2004)
7. Cai, D., He, X., Li, Z., Ma, W.Y., Wen, J.R.: Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information. In: MULTIMEDIA 2004: Proceedings of the 12th annual ACM International Conference on Multimedia, pp. 952–959. ACM Press, New York (2004)
8. Li, Z., Xu, G., Li, M., Ma, W.Y., Zhang, H.J.: Grouping WWW Image Search Results by Novel Inhomogeneous Clustering Method. In: Chen, Y.P.P. (ed.) 11th International Conference on Multi Media Modeling (MMM 2005), pp. 255–261. IEEE Computer Society, Los Alamitos (2005)
9. Gao, B., Liu, T.Y., Qin, T., Zheng, X., Cheng, Q.S., Ma, W.Y.: Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts. In: MULTIMEDIA 2005: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 112–121. ACM Press, New York (2005)

10. Gospodnetic', O., Hatcher, E.: Lucene in Action (2005)
11. Lux, M., Granitzer, M.: Retrieval of MPEG-7 Based Semantic Descriptions. In: BTW-Workshop WebDB Meets IR at the GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web, University Karlsruhe (2005)
12. J.M. (ed.): MPEG-7 Overview (version 10). Technical Report N6828, ISO/IEC JTC1/SC29/WG11 (MPEG) (October 2004), online August 17, 2007, <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>

Trans-Media Pseudo-Relevance Feedback Methods in Multimedia Retrieval

Stephane Clinchant, Jean-Michel Renders, and Gabriela Csurka

Xerox Research Centre Europe, 6 ch. de Maupertuis, 38240 Meylan, France

`FirstName.LastName@xrce.xerox.com`

<http://www.xrce.xerox.com>

Abstract. We present here some transmedia similarity measures that we recently designed by adopting some “intermediate level” fusion approaches. The main idea is to use some principles coming from pseudo-relevance feedback and, more specifically, transmedia pseudo-relevance feedback for enriching the mono-media representation of an object with features coming from the other media. One issue that arises when adopting such a strategy is to determine how to compute the mono-media similarity between an aggregate of objects coming from a first (pseudo-)feedback step and one single multimodal object. We propose two alternative ways of addressing this issue, that result in what we called the “transmedia document reranking” and “complementary feedback” methods respectively.

For the ImageCLEF - Photo Retrieval Task, it appears that mono-media retrieval performance is more or less equivalent for pure image and pure text content (around 20% MAP). Using our transmedia pseudofeedback-based similarity measures allowed us to dramatically increase the performance by $\sim 50\%$ (relative). From a cross-lingual perspective, the use of domain-specific, corpus-adapted probabilistic dictionaries seems to offer better results than the use of a broader, more general standard dictionary. With respect to the monolingual baselines, multilingual runs show a slight degradation of retrieval performance (~ 6 to 10% relative).

Keywords: hybrid retrieval, trans-media relevance feedback.

1 Introduction and Related Works

Up to now, many standard methods to tackle the problem of defining efficient trans-modal similarity measures and of solving the associated *semantic gap* use “late fusion” strategies. Basically, they rely on mono-modal analysis of multifacet objects, computing mono-modal similarities independently and then merging these mono-modal similarities by some simple aggregation operator.

In contrast to these strategies, we propose here two “intermediate level” fusion approaches. Both approaches, similarly to [1,2,3], are based on Transmedia Relevance Pseudo-Feedback, i.e. they are mixed-modality extensions of Relevance Models in which the modality of data is switched during the (pseudo-) feedback process, from image to text or text to image.

Our first approach, called Complementary Feedback (section 3.1), is similar to the approach suggested by [3]. However, while [3] uses classical text-based feedback algorithms (like Rocchio), we use a pseudo-feedback method issued from the language modelling approach to information retrieval, namely the mixture model method from Zhai and Lafferty [4] originally designed to enrich textual queries.

In our second approach, called Transmedia Document Reranking Approach (section 3.2), we do not really extract a new query, nor enrich an existing one. This second approach uses the similarity computed in the other mode as a component of feedback, in order to rerank documents. So, this is a one step retrieval, contrarily to the first one (and the related works). This method is quite general since it can be applied to any textual/visual similarities or, equivalently, with any mono-modal (textual / visual) retrieval engine. This is not the case for the other methods: [1], for instance, is based on a specific similarity model both for texts and images. Moreover, as the alternative methods require a second retrieval step, the use of a particular choice of text feedback method depends implicitly on the underlying text retrieval model. Our method is free from such dependencies, since it works on similarities as basic components.

Even if both approaches appear to be rather simpler than most alternative state-of-the-art approaches, they turned out to give superior results in the ImageClef PhotoRetrieval Track ([5]).

2 Monomedia Similarities

2.1 Cross-Entropy between Texts

Starting from a traditional bag-of-word representation of pre-processed texts (here, preprocessing includes tokenization, lemmatization, word decomposing and standard stopword removal), we adopt the language modeling approach to information retrieval and we use the (asymmetric) cross-entropy function as similarity. Particular details of this textual similarity measure are given in [6].

2.2 Fisher Vectors for Images

To compute the similarity measure between images I and J , we simply use the the L1-norm of the difference between their Fisher vectors (normalised gradient vector of the corresponding generative model, with unitary L1-norm; see details in [6,7]).

3 Cross-Media Similarities Based on Transmedia Relevance Feedback

The main idea is the following: for a given image i , consider as new features the (textual) terms of the texts associated to the most similar images (from a purely visual viewpoint). We will denote this neighbouring set as $N_{img}(i)$. Its size is fixed a priori: this is typically the top N objects returned from a retrieval system

(CBIR) or the N nearest-neighbours using some predefined visual similarity measures. Then, we can compute a new similarity with respect to any multimodal object j of the collection $\mathcal{O}$ as the textual similarity of this new representation of the image i with the textual parts of j .

There are three families of approaches to compute the mono-media similarity between an aggregate of objects $N_{img}(i)$ and one single multimodal object:

1. aggregating $N_{img}(i)$ to form a single object (typically by concatenation) and then compute standard similarity between two objects;
2. use a method of pseudo feedback algorithm (for instance Rocchio's algorithm) to extract relevant, meaningful features of an aggregate and finally use a mono-media similarity.
3. aggregating all similarity measures (assuming we can) between all possible couple of objects

Methods of families 1 and 2 involve therefore the creation of a “new single object” (in this case, a text) and a new retrieval step (this time using a text retrieval system). The third family does not.

3.1 Complementary Pseudo-Feedback

This approach, as the work presented in [3], belongs to the second family of aggregation strategies mentioned in the previous section but, contrarily to [3], our method uses the Language Modelling framework to realize the pseudo-feedback.

Recall that the fundamental problem in transmedia feedback is to define how we compute the mono-media similarity between an aggregate of objects $N_{img}(i)$ (or $N_{txt}(i)$) and one single multimodal object. The main idea here is to consider the set $N_{img}(i)$ as the “relevance concept” $\mathbf{F}$ and derive its corresponding language model (LM) θ_F . Afterwards, we can use the cross-entropy function between θ_F and the LM of the textual part of any object j in $\mathcal{O}$ as the new transmedia similarity.

We adopt the framework given by the mixture model method from Zhai and Lafferty [4] (originally designed to enrich textual queries), to derive the LM associated with $\mathbf{F}$. See [6] for practical details.

Once θ_F has been estimated, a new query LM can be obtained through interpolation:

$$\theta_{new\_query} = \alpha\theta_{old\_query} + (1 - \alpha)\theta_F \quad (1)$$

where θ_{old_query} corresponds to the LM of the textual part of the query i .

In a nearly dual way, starting from the textual part of the query, a similar scheme using $N_{txt}(i)$ can be adopted to derive a new “visual” representation (actually some generalized Fisher Vectors) of the “relevance concept”, this time relying on Rocchio's method that is more adapted to continuous feature representation.

3.2 Transmedia Document Reranking

Unlike the Complementary Feedback, the Transmedia Document Reranking approach belongs to the third family of aggregation strategies mentioned in 3.

The main idea is to define a new cross-media similarity measure by aggregating all similarity measures (assuming we can) between all possible couple of objects retrieved by the Transmedia Relevance Feedback.

More formally, if we denote by $\mathcal{T}(u)$ the text associated to multimodal object u and by $\hat{T}(i)$ the new textual representation of image i , then the new cross-media similarity measure w.r.t. the multimodal object j is:

$$\text{sim}_{\text{ImgTxt}}(i, j) = \text{sim}_{\text{txt}}(\hat{T}(i), \mathcal{T}(j)) = \sum_{d \in N_{\text{img}}(i)} \text{sim}_{\text{txt}}(\mathcal{T}(d), \mathcal{T}(j)) \quad (2)$$

where sim_{txt} is any textual similarity measure but, in a particular embodiment, we propose to use the cross-entropy function (e.g. the one based on Language Modelling, even if it is assymetric), that appears to be one of the most effective measures in purely textual information retrieval systems.

This method can be seen as a reranking method. Suppose that q is some image query; if $T(d)$ is the text of an image belonging to the initial feedback set $N_{\text{img}}(q)$, then the rank of the own neighbors of $T(d)$ in the textual sense will be increased, even if they are not so similar from a purely visual viewpoint. In particular, this allows to define a similarity between a purely image query and a simple textual object without visual counterpart.

By duality, we can define another cross-media similarity measure: for a given text i , we consider as new features the Fisher vectors of the images associated to the most similar texts (from a purely textual viewpoint) in the multimodal database. We will denote this neighbouring set as $N_{\text{txt}}(i)$. If we denote by $\mathcal{I}(u)$ the image associated to multimodal object u and by $\hat{I}(i)$ the new visual representation of text i , then the new cross-media similarity measure is:

$$\text{sim}_{\text{TxtImg}}(i, j) = \text{sim}_{\text{img}}(\hat{I}(i), \mathcal{I}(j)) = \sum_{d \in N_{\text{txt}}(i)} \text{sim}_{\text{img}}(\mathcal{I}(d), \mathcal{I}(j)) \quad (3)$$

Finally, we can combine all the similarities to define a global similarity measure between two multi-modal objects i and j : for instance, using a linear combination,

$$\begin{aligned} \text{sim}_{\text{glob}}(i, j) = & \lambda_1 \text{sim}_{\text{txt}}(\mathcal{T}(i), \mathcal{T}(j)) + \lambda_2 \text{sim}_{\text{img}}(\mathcal{I}(i), \mathcal{I}(j)) \\ & + \lambda_3 \text{sim}_{\text{ImgTxt}}(i, j) + \lambda_4 \text{sim}_{\text{TxtImg}}(i, j) \end{aligned}$$

In one embodiment, we use a simple weighted averages of these similarities, and optimize the weights through the use of a labelled/annotated training set.

The main advantage of this method is that, using an aggregation strategy that belongs to family **3**, it does not require any further retrieval step. Furthermore, it exploits all trans-modal paths (TXT-TXT, TXT-IMG, IMG-TXT and IMG-IMG) and combines them. Finally, we can pre-compute the monomodal similarities (textual and visual) between all pairs of objects in the multimedia reference repository, as these computations are independent from the objects in the run-time application; once stored, these values can be re-injected into the translingual similarity equations at run-time, greatly reducing the computation time.

3.3 Experimental Results

Table 1 shows the name of our ImageCLEF runs and the corresponding mean average precision measures. For a description of the task, the corpus and the queries, refer to [5].

Table 1. Official Runs

Run	Txt	Img	CF1	CF2	CF3	TR1	TR2	TR3
MAP	0.21	0.189	0.317	0.29	0.278	0.28	0.276	0.30

Below is a detailed description of all the methods we used for the runs:

- **Txt:** This run was a pure text run: documents were basically preprocessed and each document was enriched using Flickr database. For each term of a document, its top 20 related tags from Flickr were added to the document (see details in [6]). Then, a unigram language model for each document is estimated, giving more weight to the original document terms. An additional step of pseudo-relevance feedback using the method explained in [4] is then performed.
- **Img:** This run is a pure image run: it uses Fisher Kernel metric to define the image similarity. As a query encompasses 3 visual sub-queries, we have to combine the similarity score with respect to these 3 subqueries. To this aim, the result lists from the image sub-queries are renormalized (by subtracting the mean and dividing by the standard deviation) and merged by simple sum.
- **CF1:** This run uses both texts and images: it starts from query images only, to determine the relevance set $N_{img}(i)$ for each query i and then implements the “the complementary (intermedia) feedback” described in section 3.1. The size of the neighbouring set is 15. Referring to the notation of section 3.1, the value of α is 0.5.
- **CF2:** This runs works with the same principle as the previous run *CF1*. The main difference is that (target) english documents have been enriched with Flickr and that the initial query — in German — was translated by multiplying its “Language Model” by the probabilistic translation matrix extracted from the (small) parallel part of the corpus. Otherwise, it uses the same parameters as previously.
- **CF3:** This run uses the same process as in *CF1*, except that it uses english queries to search for German annotations. English queries are translated with the probabilistic translation matrix extracted from the (small) parallel part of the corpus and the translated queries follow the same process as in *CF1* but with different parameter : the size of the neighbouring set is 10, while the value of α is 0.7.
- **TR1:** This run uses both texts and images: it starts from query images only, to determine $N_{img}(i)$ for each query i (as in the previous run above) and then implements the method Transmedia Reranking method described in section 3.2. The size of the neighbouring set is 5.

- **TR2:** It is basically the same algorithm as the preceding run *TR1*, except that the textual part of the data (annotations) is enriched with Flickr tags.
- **TR3:** This run uses the TR algorithm as in *TR1* but, we merge the result lists from *TR1* and from the purely text queries (*Txt*), by summing the relevance scores after normalisation (by subtracting the mean and dividing by the standard deviation for each list).

3.4 Topic-Based Analysis of Results

In order to better understand the possible correlations between the different methods and/or the systematic superiority of some of them, Figure 1 compares the Average Precisions for each pair of methods and for each topic. Methods are: text-only (TXT), image-only (IMG), our best Complementary Feedback (CF) and our best Transmedia Reranking (TR) approaches.

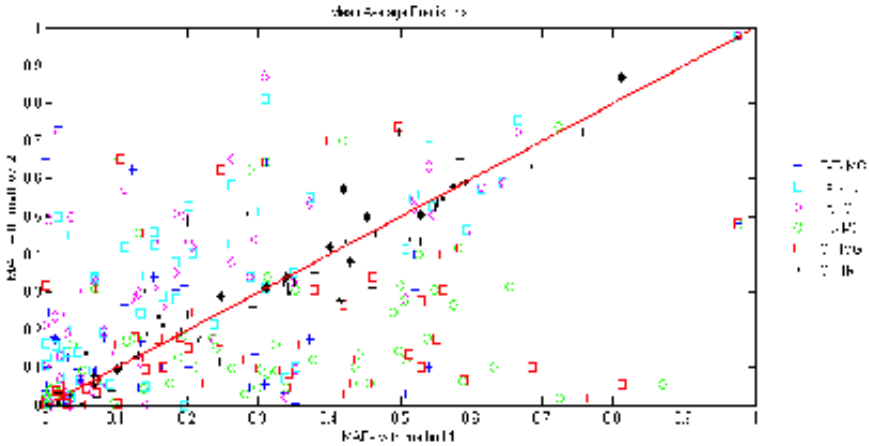


Fig. 1. Average Precision values per topic, for six pairs of methods

A deeper analysis of the individual topics leads to the following conclusions:

- From a purely visual aspect, search performance is better when the example images of the query are similar between themselves; search results degrade in the opposite case. See examples in Figure 2.
- The combination between text and image works better if the text query is complementary with respect to the visual information (see for instance left column of Figure 3).
- The combination does not perform well when either one of the media works very badly, especially the image, which is not surprising as the images were used for transmedia pseudo-relevance feedback (e.g. topics 3 and 32).
- There were also examples in which multi-media retrieval performance was poor, while individual mono-media retrieval worked not too bad (e.g. right



Fig. 2. Left column: query images from topics for which the retrieval worked best: (a) 22 – “tennis player during rally”, (b) 55 – “drawings in Peruvian desert” and (c) 51 – “photos of goddaughters from Brasil”. **Right column:** query images from topics for which the retrieval worked worst: (d) 9 – “tourist accomodation near Lake Titicaca”, (e) 10 – “destinations in Venezuela” and (f) 39 – “people in bad weather”.



Fig. 3. Left column: Query images from topics with best hybrid combinations: (a) 21 – “accomodations provided by host families”, (b) 44 – “mountains in mainland Australia” and (c) 48 – “vehicle in South Korea”. **Right column:** query images from topics with worst hybrid combinations: (d) 38 – “Machu Picchu and Huayna Picchu in bad weather”, (e) 11 – “black and white photos from Russia” and (f) 3 – “religious statue in the foreground”.

column of 3). The reason might be that the retrieved images were incorrectly reranked based on their textual similarity with the query text. For example, for topic 38, non relevant images of Machu Picchu and Huayna Picchu (because not taken under showing bad weather condition) got better ranking,

with the effect of decreasing the precision (P20 falling down from 0.7 to 0.21 (for TR) and to 0.3 (for CF)).

4 Conclusion

With a slightly annotated corpus of images, also characterised by an abstraction level in the textual description that is significantly different from the one used in the queries, it appears that mono-media retrieval performance is more or less equivalent for pure image and pure text content (around 20% MAP). Using our transmedia pseudofeedback-based similarity measures allowed us to dramatically increase the performance by $\sim 50\%$ (relative). Trying to model the textual “relevance concept” present in the top ranked documents issued from a first (purely visual) retrieval and combining this with the textual part of the original query turns out to be the best strategy, being slightly superior to our transmedia document reranking method. From a cross-lingual perspective, the use of domain-specific, corpus-adapted probabilistic dictionaries seems to offer better results than the use of a broader, more general standard dictionary. With respect to the monolingual baseline, multilingual runs show a slight degradation of retrieval performance (~ 6 to 10% relative).

Acknowledgments. This work was partly funded by the French Government under the *Infomagic* project, part of the Pole CAP DIGITAL (IMVN) de Paris, Ile-de-France. The authors also want to thank Florent Perronin for his greatly appreciated help in applying some of the Generic Visual Categorizer components in our experiments.

References

1. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: NIPS (2003)
2. Chang, Y.C., Chen, H.H.: Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. In: Working Notes of the CLEF Workshop, Alicante, Spain (2006)
3. Maillot, N., Chevallet, J.-P., Valea, V., Lim, J.H.: Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In: Working Notes of the CLEF Workshop, Alicante, Spain (2006)
4. Zhai, C., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: CIKM (2001)
5. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF-photo 2007 photographic retrieval task. In: Working Notes of the CLEF Workshop, Budapest, Hungary (2007)
6. Clinchant, S., Renders, J.-M., Csurka, G.: Xrce’s participation to ImageCLEF 2007. In: Working Notes of the CLEF Workshop, Budapest, Hungary (2007)
7. Perronin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)

Cue Integration for Medical Image Annotation

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo

IDIAP Research Institute,
Centre Du Parc, Av. des Pres-Beudin 20,
P. O. Box 592, CH-1920 Martigny, Switzerland
{`tommasi,forabona,bcaputo`}@idiap.ch

Abstract. This paper presents the algorithms and results of our participation to the image annotation task of ImageCLEFmed 2007. We proposed a multi-cue approach where images are represented both by global and local descriptors. These cues are combined following two SVM-based strategies. The first algorithm, called Discriminative Accumulation Scheme (DAS), trains an SVM for each feature, and considers as output of each classifier the distance from the separating hyperplane. The final decision is taken on a linear combination of these distances. The second algorithm, that we call Multi Cue Kernel (MCK), uses a new Mercer kernel which can accept as input different features while keeping them separated. The DAS algorithm obtained a score of 29.9, which ranked fifth among all submissions. The MCK algorithm with the one-vs-all and with the one-vs-one multiclass extensions of SVM scored respectively 26.85 and 27.54. These runs ranked first and second among all submissions.

1 Introduction

The amount of medical image data produced nowadays is constantly growing, with average-sized radiology departments producing several tera-bytes of data annually. The cost of manually annotating these images is very high and, when done manually, prone to errors [1]. This calls for automatic annotation algorithms able to perform the task reliably. The ImageCLEFmed annotation task in 2007 has provided participants with 11000 training and development data, spread across 116 classes [2]. State of the art approaches used texture-based descriptors as features and discriminative algorithms, mainly SVMs, for the classification step [3,4]. Local and global features, have been used separately or combined together in multi-cue approaches with disappointing results [3,5]. Still, years of research on visual recognition showed clearly that multiple-cue methods outperform single-feature approaches, provided that the features are complementary.

This paper describes a multi-cue strategy for biomedical image classification. We used raw pixels as global descriptors and SIFT features as local descriptors. The two feature types were combined together using two different SVM-based integration schemes. The first is the Discriminative Accumulation Scheme (DAS), proposed first in [6]. For each feature type, an SVM is trained and its output

consists of the distance from the separating hyperplane. Then, the decision function is built as a linear combination of the distances, with weighting coefficients determined via cross validation. We submitted a run using this method that ranked fifth among all submissions. The second integration scheme consists in designing a new Mercer kernel, able to take as input different feature types for each image data. We call it Multi Cue Kernel (MCK); the main advantage of this approach is that features are selected and weighted during the SVM training, thus the final solution is optimal as it minimizes the structural risk. We submitted two runs using this algorithm, the first using the one-vs-all multiclass extension of SVM; the second using instead the one-vs-one extension. These two runs ranked first and second among all submissions. These results overall confirm the effectiveness of using multiple cues for automatic image annotation.

The rest of the paper is organized as follows: section 2 describes the two types of feature descriptors we used at the single cue stage. Section 3 gives details on the two alternative SVM-based cue integration approaches. Section 4 reports the experimental procedure adopted and the results obtained, with a detailed discussion on the performance of each algorithm. The paper concludes with a summary discussion.

2 Single Cue Image Annotation

The aim of the automatic image annotation task is to classify images into a set of classes, according to the IRMA code [7]. The labels are hierarchical therefore, errors in the annotation are counted depending on the level at which the error is done and on the number of possible choices. For each image the error ranges from 0 to 1, respectively if the image is correctly classified or if the predicted label is completely wrong. The strategy we propose is to extract a set of features from each image (section 2.1) and to use then a Support Vector Machine (SVM) to classify the images (section 2.2).

2.1 Feature Extraction

We chose two types of features, local and global, with the aim to extract different informations.

Local Features. We explored the idea of “bag of words”, a common concept in many state of the art approaches to visual recognition. The basic idea is that it is possible to transform the images into a set of prespecified visual words, and to classify the images using the statistics of appearance of each word as feature vectors. To build the visual vocabulary, we used SIFT features [8], computed around interest points detected via random sampling [9]. With respect to the classic SIFT implementation, we removed the rotational invariance and the scale invariance by extracting the SIFT at only one orientation and at one octave, the one that obtained the best classification performance. To keep the complexity of the description of each image low and at the same time retain as much information as possible, we matched each extracted SIFT with a number of template

SIFTs. These template SIFTs form our vocabulary of visual words. It is built using a standard K-means algorithm, with K equal to 500, on a random collection of SIFTs extracted from the training images. Various sizes of vocabulary were tested with no significant differences, so we have chosen the smaller one with good recognition performances. Note that in this phase also testing images can be used, because the process is not using the labels and it is unsupervised. At this point each image could be described with the raw counts of each visual word. To add some kind of spatial information to our features we divided the images in four subimages, collecting the histograms separately for each subimage. In this way the dimension of the input space is multiplied by four, but in our tests we gained about 3% of classification performances. We have extracted 1500 SIFT in each subimage: such dense sampling adds robustness to the histograms. See Figures 1 for an example.

Global Features. We chose the simplest possible global description method: the raw pixels. The images were resized to 32x32 pixels, regardless of the original dimension, and normalized to have sum equal to one, then the 1024 raw pixels values were used as input features. This approach is at the same time a baseline for the classification system and a useful “companion” method to boost the performance of the SIFT based classifier (see section 2.2).

2.2 Classification

For the classification step we used an SVM with an exponential χ^2 as kernel, for both the local and global approaches:

$$K(X, Y) = \exp \left(-\gamma \sum_{i=1}^N \frac{(X_i - Y_i)^2}{X_i + Y_i} \right). \quad (1)$$

The parameter γ was tuned through cross-validation (see section 4). This kernel has been successfully applied for histogram comparison and it has been demonstrated to be positive definite [10], thus it is a valid kernel.

3 Multi Cue Annotation

Due to the fundamental difference in how local and global features are computed it is reasonable to suppose that the two representations provide different kinds of information. Thus, we expect that by combining them through an integration scheme, we should achieve a higher classification performance and a higher robustness. In the rest of the section we describe the two alternative integration schemes we used. The first, the Discriminative Accumulation Scheme (DAS, [6]), is a high-level integration scheme, meaning that each single cue first generate a set of hypotheses on the correct label of the test image, and then those hypotheses are combined together so to obtain a final output. This method is described in section 3.1. The second, the Multi Cue Kernel (MCK), is a mid-level integration scheme, meaning that the different features descriptors are kept separated

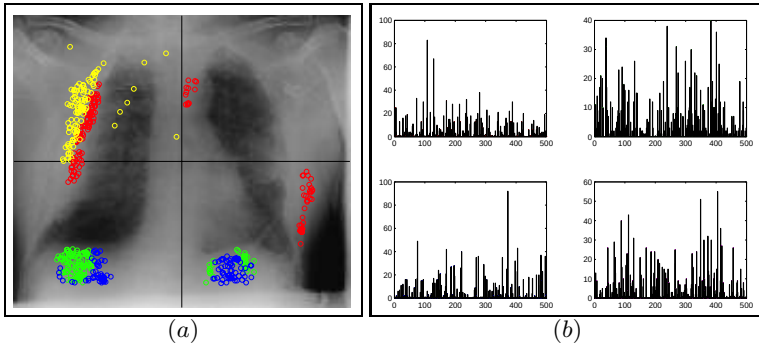


Fig. 1. (a) The four most present visual words in the image are drawn, each with a different color. and (b) total counts of the visual words in the 4 subimages.

but they are combined in a single classifier generating the final hypothesis. This algorithm is described in section 3.2.

3.1 Discriminative Accumulation Scheme

The Discriminative Accumulation Scheme is an integration scheme for multiple cues that does not neglect any cue contribution. Its main idea is that information from different cues can be summed together.

Suppose we are given M object classes and for each class, a set of N_j training images $\{I_i^j\}_{i=1}^{N_j}$, $j = 1, \dots, M$. For each image, we extract a set of P different cues so that for an object j we have P new training sets. For each we train an SVM. Kernel functions may differ from cue to cue and model parameters can be estimated during the training step via cross validation. Given a test image $\hat{I}$ and assuming $M \geq 2$, for each single-cue SVM we compute the distance from the separating hyperplane $D_j(p)$, $p = 1 \dots P$: After collecting all the distances $\{D_j(p)\}_{p=1}^P$ for all the M objects and the P cues, we classify the image $\hat{I}$ using the linear combination:

$$j^* = \underset{j=1}{\operatorname{argmax}}^M \left\{ \sum_{p=1}^P a_p D_{j(p)} \right\}, \quad \sum_{p=1}^P a_p = 1. \quad (2)$$

The coefficients $\{a_p\}_{p=1}^P$ are evaluated via cross validation during the training step.

3.2 Multi Cue Kernel

DAS can be defined a high-level integration scheme, as fusion is performed as a post-processing step after the single-cue classification stage. As an alternative, we developed a mid-level integrating scheme based on multi-class SVM with a Multi Cue Kernel K_{MC} . This new kernel combines different features extracted

form images; it is a Mercer kernel, as positively weighted linear combination of Mercer kernels are Mercer kernels themselves [11]:

$$K_{MC}(\{T_p(I_i)\}_p, \{T_p(I)\}_p) = \sum_{p=1}^P a_p K_p(T_p(I_i), T_p(I)), \quad \sum_{p=1}^P a_p = 1. \quad (3)$$

In this way it is possible to perform only one classification step, identifying the best weighting factors a_p while optimizing the other kernel parameters. Another advantage of this approach is that it makes it possible to work both with one-vs-all and one-vs-one SVM extensions to the multiclass problem.

4 Experiments

Our experiments started evaluating the performance of local and global features separately. Even if the original dataset was divided in training, validation and testing sets, we decided to merge them together and extract 5 random and disjoint train/test splits of 10000/1000 images using the cross validation technique for the parameters selection. We considered as the best parameters the ones giving the best average score on the 5 splits. Note that, according to the method used for the score evaluation, the best average score is not necessary the best recognition rate. Besides obtaining the optimal parameters, these experiments showed that the SIFT features outperform the raw pixel ones, as it was predictable.

Then we adopted the same experimental setup for DAS and MCK. In particular in DAS we used the best parameters of the previous step, so we only searched the best weights for cue integration. On the other hand, for MCK we looked for the best kernel parameters and the best feature's weights at the same time. Finally we used the results of the previous phases to run our submission experiments on the 1000 unlabeled images of the challenge test set using all the 11000 images of the original dataset as training.

The ranking, name and score of our submitted runs together with the score gain respect to the best run of other participants are listed in Table 1. Our two runs based on the MCK algorithm ranked first and second among all submissions

Table 1. Ranking of our submitted runs, name, best parameters, percentage number of SVs, score, gain respect to the best run of the other participants and recognition rate

Rank	Name	a_{sift}	a_{pixel}	#SV(%)	Score	Gain	Rec. rate
1	MCK_oa	0.80	0.20	72.0	26.85	4.08	89.7%
2	MCK_oo	0.90	0.10	64.0	27.54	3.38	89.0%
3	SIFT_oo			65.2	28.73	2.20	88.4%
4	SIFT_oa			70.0	29.46	1.47	88.5%
5	DAS	0.76	0.24	82.6	29.90	1.03	88.9%
28	PIXEL_oa			75.7	68.21	-37.28	79.9%
29	PIXEL_oo			67.1	72.42	-41.48	79.2%

stating the effectiveness of using multiple cues for automatic image annotation. It is interesting to note that even if DAS has a higher recognition rate, its score is worse than that obtained using the feature SIFT alone. This could be due to the fact that when the label predicted by the global approach, the raw pixels, is wrong, the true label is far from the top of the decision ranking.

In the same table there is also a summary of the weighting parameters for the multi-cue approaches and the number of Support Vectors (SVs) obtained showed as percentage of the total number of training vectors. As we could expect, the best feature weight (see (2) and (3)) for SIFT results higher than that for raw pixels for all the integration methods. The number of SVs is a rough indicator of the difficulty of the problem. The percentage of SVs for the MCK run, using one-vs-one multiclass SVM extension (MCK_{oa}), is slightly higher than that used by the single cue SIFT_{oa}, but lower than that used by PIXEL_{oa}. For the MCK run, using one-vs-one multiclass SVM extension (MCK_{oo}), the percentage number of SVs is even lower than that of both the single cues SIFT_{oo} and PIXEL_{oo}. These results show that combining two features with the MCK algorithm can simplify the classification problem. In general we must notice that the percentage number of support vectors is over 50%. This suggests that the classification task is challenging, and therefore the generalization properties of the method might not be optimal. For MCK_{oa}, the two classification problems with the highest number of SVs are class 1121-110-213-700 (overview image, coronal posteroanterior unspecified, nose area, musculoskeletal system) vs all, and class 1121-115-710-400 (overview image, coronal posteroanterior upright, abdomen unspecified, gastrointestinal system unspecified) vs all.

Table 2. Example of images misclassified by one or both cues and correctly classified by DAS or MCK. The values correspond to the decision rank.

				
PIXEL _{oa}	11°	1°	12°	5°
SIFT _{oa}	1°	2°	2°	5°
DAS	1°	1°	1°	2°
MCK _{oa}	1°	1°	1°	1°

Table 2 shows in details some examples of classification results. The first, second and third column contain examples of images misclassified by one of the two cues but correctly classified by DAS and MCK_{oa}. The fourth column shows an example of an image misclassified by both cues and by DAS but correctly classified by MCK_{oa}. It is interesting to note that combining local and global features can be useful to recognize images even if they are compromised by the presence of prosthesis, or reference labels put on the acquisition screen.

The confusion matrices corresponding to the single-cue, discriminative accumulation and multicue kernel approach are shown as images in Figure 2. It is

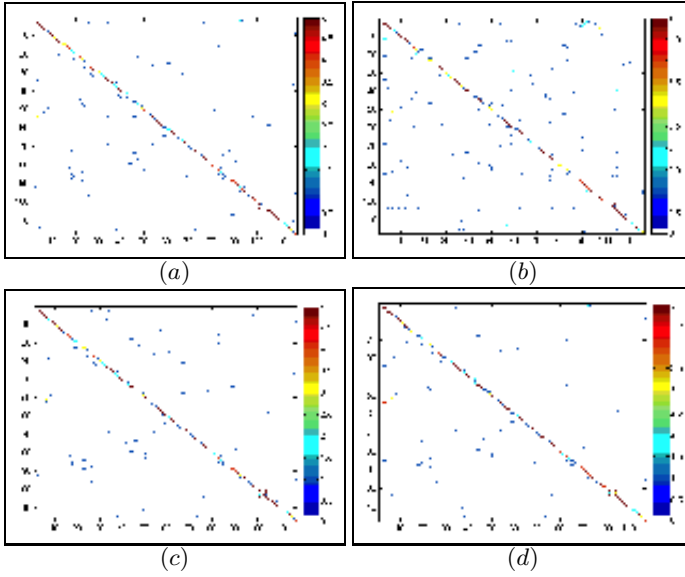


Fig. 2. These images represent the confusion matrices respectively for (a) SIFT_{oa}, (b) Pixel_{oa}, (c) DAS and (d) MCK_{oa}. To let the misclassified images stand out all the position in the matrices containing five or more images appear dark red.

clear that our methods differ principally for how the wrong images are labeled. The more the matrices present sparse values out of the diagonal and far away from it, the worse the method is. For the MCK_{oa} run the classes which contribute the most to the error score are 1123-127-500-000 confused with class 1123-110-500-000 (high beam energy, 127: coronal posteroanterior supine - 110: coronal posteroanterior unspecified chest unspecified) and class 1121-200-411-700 confused with class 1121-110-411-700 (overview image, 200: sagittal unspecified, upper extremity finger unspecified, musculoskeletal system). The class which obtains the higher benefit from the cue combination through MCK_{oa} is 1123-110-500-000, the number of correctly recognized images passes from 78 with SIFT_{oa} to 84 adding up the global (PIXEL_{oa}) information.

5 Conclusions

This paper presented a discriminative multi-cue approach to medical image annotation. We combined global and local information using two alternative fusion strategies, the Discriminative Accumulation Scheme [6] and the Multi Cue Kernel. This last method gave the best performance obtaining a score of 26.85, which ranked first among all submissions.

This work can be extended in many ways. First, we would like to use various types of local, global and shape descriptors, so to select the best features for the task. Second, our algorithm does not exploit at the moment the natural

hierarchical structure of the data, but we believe that this information is crucial. Future work will explore these directions.

Acknowledgments

This work was supported by the ToMed.IM2 project (B. C. and F. O.), under the umbrella of the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2, www.im2.ch), and by the Blanceflor Boncompagni Ludovisi foundation (T. T., www.blanceflor.se).

References

1. Güld, M.O., Kohnen, M., Keysers, D., Schubert, H., Wein, B.B., Bredno, J., Lehmann, T.M.: Quality of dicom header information for image categorization. In: *Proc of SPIE Medical Imaging*, vol. 4685, pp. 280–287 (2002)
2. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: *Working Notes of the 2007 CLEF Workshop* (2007)
3. Müller, H., Gass, T., Geissbühler, A.: Performing image classification with a frequency-based information retrieval schema for ImageCLEF 2006. In: *Working Notes of the 2006 CLEF Workshop* (2006)
4. Liu, J., Hu, Y., Li, M., Ma, W.Y.: Medical image annotation and retrieval using visual features. In: *Working Notes of the 2006 CLEF Workshop* (2006)
5. Güld, M., Thies, C., Fischer, B., Lehmann, T.: Baseline results for the imageclef 2006 medical automatic annotation task. In: *Working Notes of the 2006 CLEF Workshop* (2006)
6. Nilsback, M.E., Caputo, B.: Cue integration through discriminative accumulation. In: *Proc of CVPR* (2004)
7. Lehmann, T.M., Henning, S., Daniel, K., Michael, K., Bethold Wein, B.: The irma code for unique classification of medical images. In: *Proc of SPIE Medical Imaging*, vol. 5033, pp. 440–451 (2003)
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proc of ICCV*, vol. 2, pp. 1150–1157 (1999)
9. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3954. Springer, Heidelberg (2006)
10. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral grouping using the nyström method. *PAMI* 26(2), 214–225 (2004)
11. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge (2000)

Multiplying Concept Sources for Graph Modeling

Loic Maisonnasse¹, Eric Gaussier¹, and Jean Pierre Chevallet²

¹ LIG-UJF - 38041 Grenoble Cedex 9 France
loic.maisonnasse@imag.fr, eric.gaussier@imag.fr

² IPAL-I2R - 119613 Singapore
viscjp@i2r.a-star.edu.sg

Abstract. The main idea in this paper is to incorporate medical knowledge in the language modeling approach to information retrieval (IR). Our model makes use of the textual part of ImageCLEFmed corpus and of the medical knowledge as found in the Unified Medical Language System (UMLS) knowledge sources. The use of UMLS allows us to create a conceptual representation of each sentence in the corpus. We use these representations to create a graph model for each document. As in the standard language modeling approach, we evaluate the probability that a document graph model generates the query graph. Graphs are created from medical texts and queries, and are built for different languages, with different methods. After developing the graph model, we present our tests, which involve mixing different concepts sources (i.e. languages and methods) for the matching of the query and text graphs. Results show that using language model on concepts provides good results in IR. Multiplying the concept sources further improves the results. Lastly, using relations between concepts (provided by the graphs under consideration) improves results when only few conceptual sources are used to analyze the query.

1 Introduction

Previous ImageCLEFmed raised the interest of the use of semantic resources for IR. Indeed, some of the best performing methods from ImageCLEFmed used resources for concept extraction. As concepts can be defined as human understandable abstract notions independent from any direct material support or language, conceptual indexing solves term variation problems and is naturally multilingual. Most of the previously proposed works on concepts integrate concepts in a vector space model. We propose to improve such conceptual indexing in two ways. First we use an advanced representation of the document by using relations between concepts, thus a document is represented as a graph. Secondly we propose to extend the graph language modeling approach developed in 1 by considering that relations between terms or concepts are labeled (both syntactic and semantic relations are generally labeled; the model we present here thus addresses a common situation). This paper first presents a short overview of

the use of concepts in medical document indexing and language modeling for complex structures. Then a graph modeling approach is proposed. The different graph extraction processes used for documents and queries are then described. Finally, the different results obtained on the CLEF 2007 medical retrieval task are presented.

2 State of the Art

This section explores previous work on the use of conceptual indexing in the medical domain as well as previous work on the use of structure in language modeling.

2.1 Graphical Representations in the Medical Domain

The usefulness of concepts has been shown in the previous ImageCLEFmed tasks, where some of the best performing systems on text 2 used conceptual indexing methods based on vector space models. On TREC genomics, 3 uses the *Mesh* and *Entrez* databases to select terms from medical publications. They use terms related to concepts, by identifying these terms in document and in queries they improves the results of bag of words. They also made different experiments by adding domain specific knowledge to the query. Their results show that adding term variants provides the best improvement. If authors directly used concepts instead of terms, we believe that they would have considered variants of each concept.

Other researchers have tried to go beyond the use of concepts by exploiting relations between concepts. 4 evaluates the usefulness of UMLS concepts and semantic relations in medical IR. They first extract concepts and relations from documents and queries. To select relations in a sentence, they rely on two further assumptions: (1) interesting relations occur between interesting concepts; (2) relations are expressed by typical lexical markers such as verbs. The experiments with a vector space model show that using both concepts and relations lower the results obtained with concepts alone.

2.2 Structure Language Modeling

The language modeling approach to IR has first been proposed in 5. The basic idea is to view each document as a language sample and querying as a generative process. Even though smoothed unigram models have yielded good performance in IR, several works have investigated, within the language modeling framework, the use of more advanced representations. Works like 6 and 7 proposed to combine unigram models with bigram models. Others works, e.g. 8 or 9, incorporated syntactic dependencies in the language model. 9, for example, introduces a dependence language model for IR which integrates syntactic dependencies. This model relies on a variable L , defined as a “linkage” over query terms, which is generated from a document according to $P(L|M_d)$, where M_d represents a document model. The query is then generated given L and M_d , according to

$P(Q|L, M_d)$. In principle, the probability of the query, $P(Q|M_d)$, is to be calculated over all linkages L s, but, for efficiency reasons, the authors make the standard assumption that these linkages are dominated by a single one, the most probable one: $L = \operatorname{argmax}_L P(L|Q)$. $P(Q|M_d)$ is then formulated as:

$$P(Q|M_d) = P(L|M_d) P(Q|L, M_d) \quad (1)$$

In the case of a dependency parser, as the one used in 9, each term has exactly one governor in each linkage L , so that the above quantity can be further decomposed, leading to:

$$\log P(Q|M_d) = \log P(L|M_d) + \sum_{i=1..n} \log P(q_i|M_d) + \sum_{(i,j) \in L} MI(q_i, q_j|L, M_d) \quad (2)$$

where MI denotes the mutual information, and:

$$P(L|M_d) \propto \prod_{(i,j) \in L} \hat{P}(R|q_i, q_j) \quad (3)$$

$\hat{P}(R|q_i, q_j)$ in the above equation represents the empirical estimate of the probability that concepts q_i and q_j are related through a parse in document d . As the reader may have noticed, there is a certain ambiguity in the way the linkage L is used in this model. Consequently, this model is not completely satisfying (see 1 for a short discussion of this problem), and we rely on a different model to account for graphical structures in the language modeling approach to IR. We now describe this model, which we will refer to as the *graph model*.

3 Graph Model

We propose a graph modeling approach for IR in which each relation is labelled with one or more labels (this presentation generalizes the one in 1 and simplifies the one in 10). We assume that a semantic analysis of a query q can be represented as a graph $G_q = (C, E)$, where C is the set of terms (or concepts) in q , and E is a relation from $C \times C$ in the set of the label sets EN ($E(c_i, c_j) = \{labels\}$ if c_i and c_j are related through a relation labelled with the labels in $\{labels\}$, and $\emptyset$ otherwise). The probability that the graph of query q is generated by the model of document d can be decomposed as:

$$P(G_q|M_d) = P(C|M_d) P(E|C, M_d) \quad (4)$$

Assuming that, conditioned on M_d , query concepts are independent of one another (a standard assumption in the language model), and that, conditioned on M_d and C , edges are independent of one another (again a standard assumption), we can write:

$$P(C|M_d) = \prod_{c_i \in C} P(c_i|M_d) \quad (5)$$

$$P(E|C, M_d) = \prod_{(i,j)} P(E(q_i, q_j)|C, M_d) \quad (6)$$

Equation 5 corresponds to the standard language model (potentially applied to concepts), and equation 6 carries the contribution of edges. The quantity $P(c_i|M_d)$ of equation 5 is computed through simple Jelinek-Mercer smoothing, using a smoothing parameter λ_u .

The quantities $P(E(q_i, q_j)|C, M_d)$ of equation 6 can be decomposed as:

$$P(E(q_i, q_j)|C, M_d) = \prod_{label \in E(q_i, q_j)} P(R(q_i, q_j, label)|q_i, q_j, M_d) \quad (7)$$

where $R(q_i, q_j, label)$ indicates that there is a relation between q_i and q_j , the label set of which contains $label$.

An edge probability is thus equal to the product of the corresponding single-label relations. Following standard practice in language modeling, one can furthermore “smooth” this estimate by adding a contribution from the collection. This results in:

$$P(R(c_i, c_j, label)|C, M_d) = (1 - \lambda_e) \frac{D(c_i, c_j, label)}{D(c_i, c_j)} + \lambda_e \frac{C(c_i, c_j, label)}{C(c_i, c_j)} \quad (8)$$

where $D(c_i, c_j, label)$ ($C(c_i, c_j, label)$) is the number of times c_i and c_j are linked with a relation labeled $label$ in the document (collection). $D(c_i, c_j)$ ($C(c_i, c_j)$) is the number of times c_i and c_j are observed together in the document.

The above model can be applied to any graphical representation of queries and documents, and relies on only two terms, which are easy to estimate. We now show how this model behaves experimentally.

4 Graph Extractions

UMLS is a good candidate as a knowledge source for medical text indexing. It is more than a terminology because it describes terms with associated concepts. But it is not an ontology, as there is no formal description of concepts. Nevertheless, the large set of terms and term variants in UMLS (more than 1 million concepts associated with 5.5 million of terms) restricted to medical domain, allows one to build on top of it a full scale conceptual indexing system. In UMLS, all concepts are assigned to at least one semantic type from the Semantic Network. This enables to detect general semantic relation between concepts. With UMLS, we produce graphs from a text in two steps: by first detecting concept, and by detecting relations between detected concepts of a same sentence. These two steps are detailed in 10. Results of a graph extraction for a sentence can be viewed on figure 1.

To cover all collection languages, we use three variations of concept detection. We therefore obtained three graph extraction methods:

- (1) uses MetaMap11, for English only.
- (2) uses a term mapping with MiniPar, for English only.
- (3) uses a term mapping with TreeTagger, over all languages.



Fig. 1. Graph produced for *Show me chest CT images with emphysema*

We compute the frequency $D(c_i, c_j, name)$ (resp. $C(c_i, c_j, name)$) as the number of times c_i and c_j are linked in sentences of the document (resp. collection). $D(c_i, c_j)$ (resp. $C(c_i, c_j)$) corresponds to the number of times c_i and c_j appear in the same sentence of a document (resp. collection). Hence, because relations are built between concepts co-occurring in the same sentence, the probability of a relation in a document will be 1 if the two concepts appear in the same sentence and if the concepts are linked in the semantic network. It will be 0 otherwise.

5 Evaluation

We show here the results obtained with the previous methods on the corpus CLEFmed 2007 12.

5.1 Collection Analysis

We analyse the English part of the collection with MetaMap(1) and the French and the German parts with TreeTagger(3).

For queries, we group analyses made in different languages (and using different tools). A query is therefore defined as a set of graphs $Q = \{G_q\}$. The probability of a query assuming a document graph model is obtained by the product of the probability of each query graph.

$$P(Q = \{G_q\} | M_g) = \prod_{G_q} P(G_q | M_d) \quad (9)$$

We propose to group the analyses as follows:

- (E) one English graph produced by (1)
- (E_Mix) English graphs produced by (1)(2)(3)
- (EFG) English graph produced by (1) with French and German graphs produced by (3)
- (EFG_Mix) English graphs produced by (1)(2)(3) with French and German graphs produced by (3) (thus leading to 5 different graphs).

Table 1. Results for mean average precision (MAP) and precision at five documents (P@5)

	unigram model				graph model				
	2005-2006		2007		2005-2006		2007		
	λ_u	text	image	image	λ_u	λ_e	text	image	image
MAP									
E	0.2	0.2468	0.2284	0.3131	0.2	0.9	0.2463	0.2277	0.3271
E_Mix	0.1	0.2610	0.2359	0.3376	0.1	0.9	0.2620	0.2363	0.3377
EFG	0.1	0.2547	0.2274	0.3269	0.1	0.9	0.2556	0.2313	0.3345
EFG_Mix	0.1	0.2673	0.2395	0.3538	0.1	0.9	0.2670	0.2394	0.3536
P@5									
E	0.2	0.4618	0.4436	0.3733	0.2	0.9	0.4582	0.4400	0.4133
E_Mix	0.1	0.4727	0.4582	0.3667	0.1	0.8	0.4800	0.4582	0.3667
EFG	0.2	0.4582	0.4364	0.4467	0.1	0.8	0.4618	0.4473	0.4867
EFG_Mix	0.1	0.4836	0.4691	0.4200	0.1	0.8	0.4909	0.4655	0.4200

5.2 Global Results

We first evaluate our system on the two previous years of CLEFmed, and select the best performing methods at the textual level. At this level, we consider a textual annotation relevant if one of its associated images is relevant at the image level. Table 1 shows the results obtained on CLEF 2007 for the different collections, with the best parameters evaluated on the textual part of CLEFmed 2005 and 2006. We evaluate the results with mean average precision (MAP), since it gives an overview of results, and with precision at 5 documents (P@5), since this measure shows system precision on first results. Results show that the best performing method, for MAP, is the one that uses all concept sources (EFG_mix), this results is the best for CLEF 2007. Using different concept sources for the query improves the overall results of IR. Such a method helps finding all query concepts and improves the recall. But for precision at five documents, best results are obtained with EFG that use one concept source per language. Using only the three languages provides the best concepts. Adding other concept sources may add some false concepts that lower the precision. For graph indexing, MAP results show a similar behaviour to concepts alone. The only difference is on EFG where relation results are better than concept alone, on MAP and P@5. The results obtained on the MAP indicate that considering less alternative graphs for the English queries leads to higher precision at the expense of recall), and that relation help for high precision tasks.

5.3 Results by Query Types

CLEF queries are divided in three types (visual, textual and mixed). We evaluate the impact of our model, with the previously selected parameters, depending on query types. The results presented in table 2 show that adding concept sources

Table 2. Concepts and relations statistics and results depending on query type

	concepts		relations		unigram model		graph model	
	detected	distinct	detected	distinct	MAP	P@5	MAP	P@5
Visual								
E	19	46	213	210	0.2462	0.3600	0.2460	0.3600
EFG	62	50	221	218	0.2280	0.3400	0.2250	0.3400
EFG_Mix	117	51	335	218	0.2203	0.2400	0.2190	0.2400
Mixed								
E	43	39	282	276	0.2600	0.1600	0.2945	0.2400
EFG	60	43	290	282	0.3189	0.3800	0.3366	0.4800
EFG_Mix	115	43	429	282	0.3471	0.3800	0.3068	0.2200
Textual								
E	55	46	528	512	0.4327	0.6000	0.4409	0.6200
EFG	61	47	529	513	0.4330	0.6000	0.4419	0.6400
EFG_Mix	122	49	786	514	0.4939	0.6400	0.4873	0.6400

mainly improves textual and mixed results, but decreases the performance of the system for visual queries. Visual queries are made of simple concepts easily identified in English; adding other concept sources in this case seems to merely increase the noise level.

In particular, the unigram model works well for textual queries, ie queries focussing on one or two concepts. For theses queries, adding concept sources improves the recall without decreasing the precision. A detailed analysis of the 10 2007 textual queries shows that 6 of them have a P@5 equal to 1 with our system, and that 4 provide the best MAP at CLEF 2007 for most runs (over all participants). This means that textual queries are relatively easy to handle compared to the other types. Our model does not work as well for visual queries, inasmuch as using multiple concept sources for such queries decreases the results. Indeed, for the method EFG_Mix, 6 of the 10 visual queries have a P@5 of 0. Most of the time, the concept(s) expressed in a visual query corresponds to a precise modality, as *cardiac MRI*. Our investigation leads us to think that such concepts are usually implicit in the written documents, and thus must be extracted from the image (a difficult task we have not addressed here).

On both textual and image queries, using relations has no impact: extracting the main (or the only) concept is enough. On mixed queries, though, relations do have an impact, which we believe is due to the fact that these queries are complex and often comprise more than one concept (contrary to the other types). However, as the parameters of our model are estimated on all query types, the weight of the relation contribution will be small (due to textual and image queries). In fact, it may be too small to observe all the impact relations may have on mixed queries. One solution to this problem would be to adapt the model to the different query types. We plan to do so once we will have enough query of each type.

6 Conclusion

We presented here a framework for using semantic resources in the medical domain by describing a method for building graphical representations of documents, and proposing a graph modeling approach for IR. Within this framework, we evaluated the impact of using multiple concept sources for analysing queries. Our results show that graph indexing can be useful for improving the precision at the top of the list (P@5), and that multiplying concept sources improves the overall results (MAP and P@5) of IR. In our experiments, we only analysed queries. In future work, we intend to evaluate the use of multiple concept sources for analysing documents as well. Furthermore, the relation extraction method we have relied on is relatively simple. We believe that using a more sophisticated method could further improve our results.

References

- Maisonnasse, L., Gaussier, E., Chevallet, J.-P.: Revisiting the dependence language model for information retrieval. In: *Research and Development in Information Retrieval* (2007)
- Lacoste, C., Chevallet, J.-P., Lim, J.-H., Wei, X., Raccoceanu, D., Le, T.-H.D., Teodorescu, R., Vuillenemot, N.: Ipal knowledge-based medical image retrieval in image-clefmed 2006. In: *Working Notes for the CLEF 2006 Workshop*, 20-22 September, Alicante, Spain, (2006)
- Zhou, W., Yu, C., Smalheiser, N., Torvik, V., Hong, J.: Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: *Research and Development in Information Retrieval* (2007)
- Vintar, S., Buitelaar, P., Volk, M.: Relations in concept-based cross-language medical information retrieval. In: *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining (ATEM)* (2003)
- Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: *Research and Development in Information Retrieval* (1998)
- Srikanth, M., Srikanth, R.: Biterm language models for document retrieval. In: *Research and Development in Information Retrieval* (2002)
- Song, F., Croft, W.B.: A general language model for information retrieval. In: *CIKM 1999: Proceedings of the eighth international conference on Information and knowledge management*, pp. 316–321. ACM Press, New York (1999)
- Lee, C., Lee, G.G., Jang, M.G.: Dependency structure language model for information retrieval. *ETRI journal* (2006)
- Gao, J., Nie, J.Y., Wu, G., Cao, G.: Dependence language model for information retrieval. In: *Research and Development in Information Retrieval* (2004)
- Maisonnasse, L., Gaussier, E., Chevallet, J.P.: Multiplying concept sources for graph modeling. In: *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 19-21 September (2007)
- Aronson, A.: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: *Proc AMIA 2001*, pp. 17–21 (2001)
- Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary (2007)

MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval

Julio Villena-Román^{1,3}, Sara Lana-Serrano^{2,3}, and José Carlos González-Cristóbal^{2,3}

¹ Universidad Carlos III de Madrid

² Universidad Politécnica de Madrid

³ DAEDALUS - Data, Decisions and Language, S.A.

jvillena@it.uc3m.es, slana@diatel.upm.es,
josecarlos.gonzalez@upm.es

Abstract. This paper describes the participation of MIRACLE research consortium at the ImageCLEF Medical Image Retrieval task of ImageCLEF 2007. For this campaign, our challenge was to research on different merging strategies, i.e. methods of combination of textual and visual retrieval techniques. We have focused on the idea of performing all possible combinations of well-known textual and visual techniques in order to find which ones offer the best results in terms of MAP and analyze if the combined results may improve the individual ones.

Keywords: Image retrieval, domain-specific vocabulary, thesaurus, linguistic engineering, information retrieval, indexing.

1 Introduction

MIRACLE is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a small/medium size enterprise founded as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain.

This paper describes our participation [1] in the ImageCLEFmed task of ImageCLEF 2007 [2], whose goal is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as text.

2 System and Experiment Description

The system is built up from three different modules. The first one is the **textual (text-based) retrieval module**, which indexes case descriptions in order to look for the most relevant ones to the text of the topic. The system consists of a set of different basic components organized in two categories: resources and tools for medical-specific vocabulary analysis and linguistic tools for textual analysis and retrieval.

Instead of using raw terms, the textual information of both topics and documents is parsed and tagged to unify all terms into concepts of medical entities. Thus, concept identifiers [3] are used instead of terms in the text-based process of information retrieval. For this purpose, a terminological dictionary was created by using a subset of the Unified Medical Language System (UMLS) metathesaurus [4] and incorporating terms in English, Spanish, French and German (the languages involved in the task). This dictionary contains 4,327,255 entries matching 1,215,749 medical concepts

The baseline approach to process the document collection is based on the following steps which are executed sequentially: text extraction, medical-vocabulary recognition, tokenization, lowercase words, filtering, stemming, indexing and retrieval. Lucene [5] engine was used for all textual indexing and retrieval tasks.

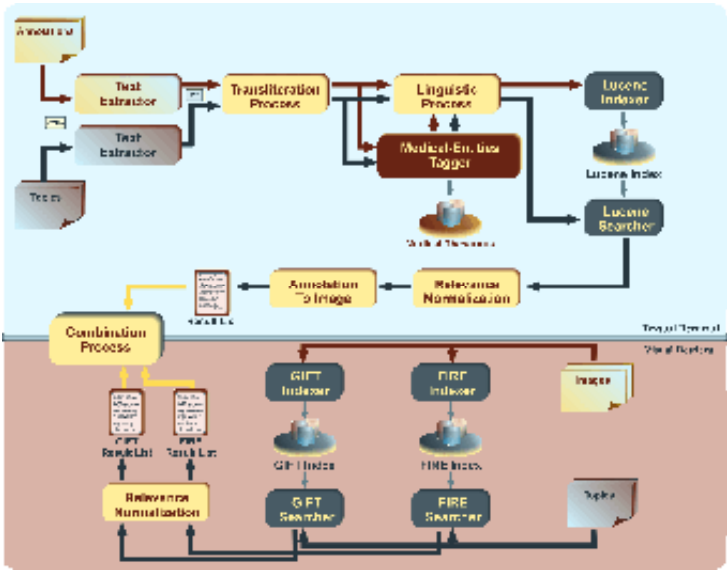


Fig. 1. Overview of the system

The second component is the **visual (content-based) retrieval module**, which provides the list of case images that are more similar to the topic ones. For this part of the system, we resorted to two publicly and freely available CBIR systems: GIFT (GNU Image Finding Tool) [6] and FIRE (Flexible Image Retrieval Engine) [7].

The main component is the **result combination module**, which uses different operators to combine the results of the two previous subsystems. The textual and image result lists are merged by applying different techniques, which are characterized by an operator and a metric for computing the relevance (score) of the result, as shown in Tables 1 and 2.

Experiments are defined by the choice of different combinations of the previously described modules, operators and score computation metrics. A wide set of experiments was submitted: 8 text-based runs covering the 3 different topic languages, 9 content-based runs (built with the combination of results from GIFT and FIRE), and also 33 mixed runs (built with the combination of textual and visual experiments).

Table 1. Combination operators

Operators	
OR	$A \cup B$
AND	$A \cap B$
LEFT	$(A \cup B) \cup (A - B)$
RIGHT	$(A \cup B) \cup (B - A)$

Table 2. Score computing metrics

Metrics	Score
max	$\max(a, b)$
min	$\min(a, b)$
avg	$\text{avg}(a, b)$
	$\max(a, b) +$
mm	$\min(a, b) * \frac{\min(a, b)}{\max(a, b) + \min(a, b)}$

3 Result Analysis

Best runs are shown in Table 3. The highest MAP of the text-based experiments is obtained by the baseline experiment in English where only stemming plus stopword removal is performed. Surprisingly for us, tagging with UMLS thesaurus has proved to be of no use with regards to the simplest strategy. We have detected some bugs in the scripts for the generation of the result sets which may have led to those results.

Table 3. Best results for textual, visual and mixed experiments

	RelRet	MAP	R-prec	P10	P30	P100
TxtENN	1842	0.3385	0.3772	0.4567	0.3578	0.293
TxtXN	1784	0.2647	0.3139	0.3367	0.3167	0.2467
TxtENT	1608	0.2480	0.2567	0.3533	0.3011	0.2227
VisG	496	0.0182	0.0371	0.0800	0.0767	0.0443
VisGFANDmm	156	0.0100	0.0256	0.0667	0.0500	0.0340
VisGFANDmax	156	0.0098	0.0252	0.0600	0.0511	0.0337
MixGENTRIGHTmin	1608	0.2480	0.2567	0.3533	0.3011	0.2227
MixGENTRIGHTmax	1648	0.2225	0.2527	0.3200	0.2856	0.2443
MixGENTRIGHTmm	1648	0.2191	0.2522	0.3267	0.2967	0.2443

Experiments using French and German languages achieve a very low precision (31% and 28% with regards to English). This result is similar to other experiments carried out in other CLEF tracks [8] and may be attributed to deficient stemming [9].

In general, MAP values of content-based experiments are very low, which reflects the complexity and difficulty of the visual-only retrieval for this task. The best value (5% of the top ranked textual experiment) is obtained with the baseline visual experiment, which just uses GIFT. From other groups' results, we feel that FIRE turns out to be better than GIFT for medical images, which could be explained because FIRE includes more features focused on grayscale images such as a higher number of histogram gray levels.

Regarding the mixed experiments, the evaluations for the experiments with the OR operator were missing in the Excel files provided by the task organizers. Many images are filtered out by the restrictive AND operator (165 instead of 532 relevant images retrieved), which the OR operator would have kept in the result list. Thus,

although the MAP of the best ranked mixed experiment is lower than the MAP of the best textual one (73%), we cannot conclude that the combination of textual and visual results with any kind of merging strategy fails to improve the precision. Note that the best ranked runs are those with the RIGHT operator, which implicitly includes an OR. In addition, the use of this operator (visual RIGHT textual) shows that textual results are preferred over visual ones (RIGHT prioritizes the second list).

Another conclusion to be drawn from these results is that the textual retrieval is the best strategy for this task. In addition, the best experiment at ImageCLEFmed 2007 [2] reaches a MAP value of 0.3538, slightly better than ours. It was also a textual-only experiment, which confirms this idea. We think that the reason is because many queries include semantic aspects such as medical diagnoses or specific details present in the image, which a purely visual retrieval cannot tackle. This issue will be considered for future participations.

Despite this difference with the best experiment, MIRACLE participation is ranked 3rd out of over 12 groups, which is indeed considered to be a very good position. It is also important to note that our early precision (P5 and P10) is the highest among all groups. Thus, relevance feedback techniques will be included next year in our system to give more importance to the first results returned by the system.

References

1. Villena-Román, J., Lana-Serrano, S., González-Cristóbal, J.C.: MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval. In: Working Notes of the Cross Language Evaluation Forum 2007, Budapest, Hungary (2007)
2. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
3. González, J.C., Villena, J., Moreno, C., Martínez, J.L.: Semiautomatic Extraction of Thesauri and Semantic Search in a Digital Image Archive. In: Integrating Technology and Culture: 10th International Conference on Electronic Publishing, ELPUB, Bansko, Bulgaria (2006)
4. U.S. National Library of Medicine. National Institutes of Health, <http://www.nlm.nih.gov/research/umls/>
5. Apache Lucene project, <http://lucene.apache.org>
6. GIFT: GNU Image-Finding Tool, <http://www.gnu.org/software/gift/>
7. FIRE: Flexible Image Retrieval System, <http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html>
8. Martínez-Fernández, J.L., Villena-Román, J., García-Serrano, A.M., González-Cristóbal, J.C.: Combining Textual and Visual Features for 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. Springer, Heidelberg (2006)
9. Martínez-Fernández, J.L., Villena-Román, J., García-Serrano, A.M., Martínez-Fernández, P.: MIRACLE team report for ImageCLEF IR in CLEF 2006. In: Proceedings of the Cross Language Evaluation Forum 2006, Alicante, Spain (2006)

MIRACLE at ImageCLEFanoT 2007: Machine Learning Experiments on Medical Image Annotation

Sara Lana-Serrano^{1,3}, Julio Villena-Román^{2,3}
José Carlos González-Cristóbal^{1,3}, and José Miguel Goñi-Menoyo¹

¹ Universidad Politécnica de Madrid

² Universidad Carlos III de Madrid

³ DAEDALUS - Data, Decisions and Language, S.A.

slana@diatel.upm.es, jvillena@it.uc3m.es

josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es

Abstract. This paper describes the participation of MIRACLE research consortium at the ImageCLEF Medical Image Annotation task of ImageCLEF 2007. Our areas of expertise do not include image analysis, thus we approach this task as a machine-learning problem, regardless of the domain. FIRE is used as a black-box algorithm to extract different groups of image features that are later used for training different classifiers based on kNN algorithm in order to predict the IRMA code. The main idea behind the definition of our experiments is to evaluate whether an axis-by-axis prediction is better than a prediction by pairs of axes or the complete code, or vice versa.

Keywords: Medical image, image annotation, classification, IRMA code, axis, learning algorithms, nearest-neighbour, machine learning.

1 Introduction

MIRACLE is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a private company founded as a spin-off of these groups and a leading company in the field of linguistic technologies in Spain. This paper describes our second participation [1] [2] in the ImageCLEF Medical Image Annotation task of ImageCLEF 2007 [3]. We approach this task as a machine learning problem, regardless of the domain, as our areas of expertise do not include image analysis research [4] and this task uses no textual information.

2 Description of Experiments

FIRE (Flexible Image Retrieval Engine) [5] [6] is a freely available content-based information retrieval system developed under the GNU General Public License that allows to perform query by example on images, using an image as the starting point for

the search process and relying entirely on the image contents. FIRE offers a wide repository of available features and distance functions. Specifically, the distribution package includes a set of scripts that extracts different types of features from the images [4], including color/gray histograms, invariant features histograms, Gabor features, global texture descriptor, Tamura features, etc.

Our approach to the task is to build different classifiers that use image features to predict the IRMA code [7]. For that purpose, all images in the training, development and testing dataset have been processed with FIRE. The extracted features have been divided into three groups, as shown in Table 1, to build the training data matrixes for the classifiers.

Table 1. Training data matrixes

Name	FIRE – Image Features	Dimension ¹
Histogram	Gray histogram and Tamura features	768
Vector	Aspect ratio, global texture descriptor and Gabor features	75
Complete	Gray histogram, Tamura features, aspect ratio, global texture descriptor and Gabor features	843

Different strategies have been evaluated, using several multiclassifiers built up with a set of specialized individual classifiers [2]:

- **IRMA Code Classifier:** single classifier that uses the image features to predict the complete IRMA code (4 axes: Technical, Direction, Anatomical and Biological).
- **IRMA Code Axis Classifier:** a two level classifier that is composed of four different classifiers that individually predict the value of each axis of the IRMA code; the prediction is the concatenation of partial solutions.
- **IRMA Code Combined Axis Classifier:** similar to the axis classifier, this one predicts the axes grouped in pairs.

All classifiers are based on the k-Nearest-Neighbour algorithm to predict the output class. After some preliminary runs, a value of k=10 was chosen. The main idea behind the definition of the experiments is to evaluate whether an axis-by-axis prediction is better than a prediction by pairs of axes or the complete code, or vice versa. In addition, the effect of applying the data normalization will be also analyzed. Table 2 shows an overview of the experiments. 30 experiments were finally submitted.

Table 2. Experiment set

Features	Prediction ²	Normalization ³
Complete Histogram Vector	Complete code Axis-by-axis Combined axis: T+A and B+D Combined axis: T+B and D+A Combined axis: T+D and A+B	NO YES

¹ Number of columns of the matrix; the number of rows is 10,000 for the training dataset and 1,000 for the development and testing dataset.

² IRMA code axes are: Technical (**T**), Direction (**D**), Anatomical (**A**) and Biological (**B**).

³ Normalized to range [0, 1].

3 Results

Results are shown in Table 3 [2]. According to the weighted error count score [8], which penalizes wrong decisions that are easy to take over wrong difficult decisions or at an early stage in the IRMA code, our best experiment is the one with data normalization that individually predicts each axis using all image features (“histogram” and “vector”). However, considering the number of correctly classified images, the best experiment is the one that uses normalized vector-based features and predicts the combined axis Technical+Direction and Anatomical+Biological.

Table 3. Evaluation of best-ranked experiments

Run ID	Error count	Well classified
MiracleAAn	158.82	497
MiracleVAn	159.45	504
MiracleAATDABn	160.25	501
MiracleAATABDn	162.18	499
MiracleVATDABn	174.99	507

On the other hand, comparing the predictions of the complete IRMA code versus the axis-by-axis predictions, the conclusion is that, regardless of the selected image features, the axis-by-axis prediction achieves more accurate results not only than the prediction of a combined pair of axes but also than the prediction of the complete code. It is interesting to observe that most groups have performed experiments focused on the prediction of the complete code.

In addition, data normalization seems to improve the predictions and vector-based features are preferred over histogram-based ones [2].

Our results were considerably worse, ompared to other groups’. The best experiment reached a score of 26.84, 17% of our own best error count. MIRACLE ranked 9th out of 10 participants in the task.

Probably different distance metrics should have been used to calculate the nearest neighbours. In particular, Mahalanobis distance, which is scale-invariant and takes into account the correlations among different variables, could have lead to better results.

However, we think that the main reason of the poor performance is the wrong choice of image features to train the classifiers. Although some feature selection experiments were carried out to reduce the high dimensions of the training data, no definite conclusion could be drawn and the complete set of features was finally used. Under these circumstances, the learning performance of the kNN algorithm is known to be worse than other algorithms’ such as SVM (Support Vector Machines), MLP (Multilayer Perceptrons) or Decision Trees. These classifiers will be considered for future participations in this task.

4 Conclusions and Future Work

The main conclusion that can be drawn from the evaluation is that, irrespective of the selected image features, the best experiments are those that predict the IRMA code

from the individual partial predictions of the 1-axis classifiers. Moreover, the predictions of combined pairs of axes are better than the predictions of the complete IRMA code. By extension, it could be concluded that the finer granularity of the classifier, the more accurate predictions are achieved. In the extreme case, the prediction may be built up from 13 classifiers, one per each character of the IRMA code. This issue will be further investigated and some experiments are already planned.

One of the toughest challenges to face when designing a classifier is the selection of the vector of features that best captures the different aspects that allow distinguishing one class from the others. Obviously, this requires an expert knowledge of the problem to be solved, which we currently lack. We are convinced that one of the weaknesses of our system is the feature selection. Therefore more effort will be invested in improving this topic for future participations.

Acknowledgements. This work has been partially supported 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; and by the Madrid's R&D Regional Plan, by means of the MAVIR project (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

1. Villena-Román, J., González-Cristóbal, J.C., Goñi-Menoyo, J.M., Martínez Fernández, J.L.: MIRACLE's Naive Approach to Medical Images Annotation. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria (2005)
2. Lana-Serrano, S., Villena-Román, J., González-Cristóbal, J.C., Goñi-Menoyo, J.M.: MIRACLE at ImageCLEFnet 2007: Machine Learning Experiments on Medical Image Annotation. In: Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (2007)
3. Deselaers, T., Müller, H., Deserno, T.: Automatic Medical Image Annotation in ImageCLEF 2007: Overview, Results, and Discussion. *Pattern Recognition Letters*, Special Issue on Medical Image Annotation in ImageCLEF 2007 (to appear, 2008)
4. Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. *Informing Science* 3(2), 63–66 (2000)
5. Deselaers, T., Keysers, D., Ney, H.: FIRE - Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 688–698. Springer, Heidelberg (2005)
6. FIRE: Flexible Image Retrieval System, <http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html>
7. IRMA project: Image Retrieval in Medical Applications, <http://www.irma-project.org/>
8. Deselaers, T., Kalpathy-Cramer, J., Müller, H., Deserno, T.: Hierarchical classification for ImageCLEF, Medical Image Annotation (2007), <http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef07/hierarchical.pdf>

Integrating MeSH Ontology to Improve Medical Information Retrieval

M.C. Díaz-Galiano, M.Á. García-Cumbreras, M.T. Martín-Valdivia,
A. Montejo-Ráez, and L.A. 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 participation in the ImageCLEFmed campaign. The SINAI research group has participated in the multilingual image retrieval subtask. The experiments accomplished are based on the integration of specific knowledge in the topics.

We have used the MeSH ontology to expand the queries. The expansion consists in searching terms from the topic query in the MeSH ontology in order to add similar terms. We have processed the set of collections using Information Gain (IG) in the same way as in ImageCLEFmed 2006.

In our experiments mixing visual and textual information we obtain better results than using only textual information. The weight of the textual information is very strong in this mixed strategy. In the experiments with a low textual weight, the use of IG improves the results obtained.

1 Introduction

This is the third participation of the SINAI research group at the ImageCLEFmed campaign. We have participated in the medical image retrieval subtask [1].

The main goal of the medical ImageCLEFmed task is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images and text. Queries are formulated with sample images and some textual description that explains the information needed. We have used the list of retrieved images by FIRE¹ [2], which was supplied by the organizers of this track.

Last year, we concentrated our efforts on manipulating the text descriptions associated with these images and mixing the results partial lists with the GIFT lists [3]. We also focused on preprocessing the collection using Information Gain (IG) in order to improve the quality of results and to automatize the tag selection process. However, this year we have concentrated on improving the queries using MeSH ontology. We have selected similar terms to the query in the ontology and we have added them to the query itself in order to expand it.

The following section explains the preprocessing of the collections. The query expansion is described in section 3. Sections 4 and 5 show the experiments

¹ <http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html>

accomplished and the results obtained. Finally, conclusions and further work are presented in section 6.

2 Preprocessing the Collection

In order to generate the textual collection we have used the `ImageCLEFmed.xml` file that links collections with their images and annotations. It has external links to the images and to the associated annotations in XML files. It contains relative paths from the root directory to all the related files.

The entire collection consists of six datasets (CASImage, Pathopic, Peir, MIR, endoscopic and MyPACS) with about 66,600 images (16,600 more than the previous year). Each subcollection is organized into cases that represent a group of related images and annotations. In every case a group of images and an optional annotation is given. Each image is part of a case and has optional associated annotations, which enclose metadata and/or a textual annotation. All the images and annotations are stored in separated files. The `ImageCLEFmed.xml` file only contains the connections between collections, cases, images and annotations.

The collection annotations are in XML format and most of them are in English. We have preprocessed the collections to generate one textual document per image [3].

We have used the IG measure to select the best XML tags in the collection. Once the document collection was generated, experiments were conducted with the Lemur² retrieval information system by applying the KL-divergence weighing scheme [4].

3 Expanding Queries with MeSH Ontology

The Medical Subject Headings (MeSH) is a thesaurus developed by the National Library of Medicine³. MeSH contains two organization files, an alphabetic list with bags of synonymous and related terms, called records, and a hierarchical organization of descriptors associated to the terms. We consider that a term is a set of words (no word sequence order), that is:

$$t = \{w_1, \dots, w_{|t|}\} \text{ where } w \text{ is a word} \quad (1)$$

We have used the bags of terms to expand the queries. A bag of terms is defined as:

$$b = \{t_1, \dots, t_{|b|}\} \quad (2)$$

Moreover, a term t exists in the query q ($t \in q$) if:

$$\forall w_i \in t, \exists w_j \in q / w_i = w_j \quad (3)$$

² <http://www.lemurproject.org/>

³ <http://www.nlm.nih.gov/mesh/>

Therefore, if all the words of a term are in the query, we generate a new expanded query by adding all its bag of terms.

$$q \text{ is expanded with } b \text{ if } \exists t \in b/t \in q \quad (4)$$

In order to compare the words of a particular term to those of the query, all the words are put in lowercase and no stopwords removal is applied. So as reduce the number of terms that could expand the query, we have only used those that are in A, C or E categories of MeSH (A: Anatomy, C: Diseases, E: Analytical, Diagnostic and Therapeutic Techniques and Equipment) [5]. Figure 1 shows an example of query expansion, with two terms found in the query and their bags of terms.

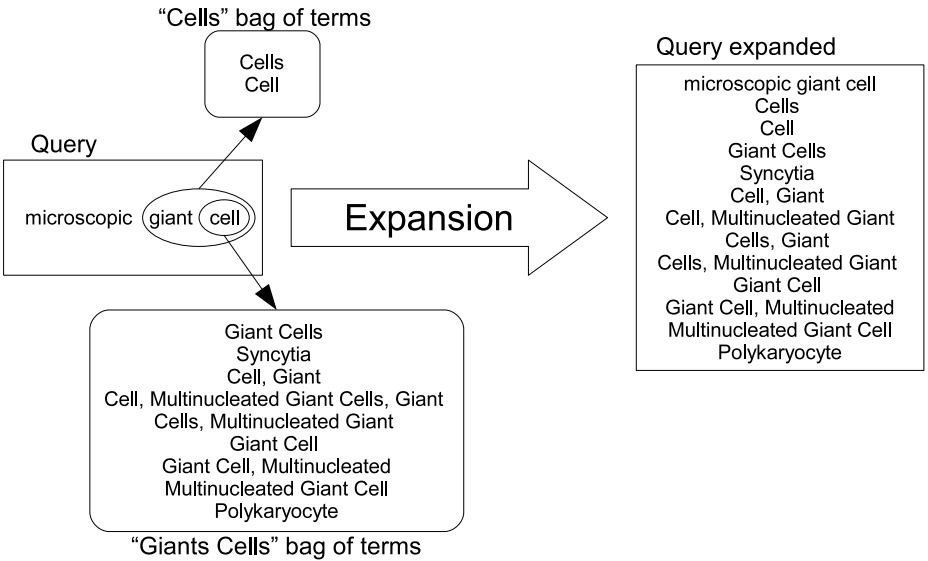


Fig. 1. Example of query expansion

4 Experiment Description

Our main objective is to investigate the effectiveness of the query expansion together with tag filtering using IG in the text collection. We have carried out these experiments using a corpus with 20%, 30%, 40%, 50% and 60% of the tags in the 2007 collection, because these settings led to the best results on the 2006 corpus.

Finally, the expanded textual list and the FIRE list are merged in order to obtain one final list (FL) with relevant images ranked by relevance. The merging process was done by giving different importance to the visual (VL) and textual lists (TL):

$$FL = TL * \alpha + VL * (1 - \alpha) \quad (5)$$

We have submitted runs with α set to 0.5, 0.6, 0.7 and 0.8.

The baseline experiments contain the 100% of the tags. To compare the mixed results we have accomplished experiments with an $\alpha = 1$.

5 Results and Discussion

The total number of runs submitted to ImageCLEFmed 2007 was 30: 6 for textual retrieval ($\alpha=1.0$), included textual baseline, and 24 for mixed retrieval. Table 1 shows the main average precision (MAP) of our runs. In this table we can observe that the best result obtained by our system is that of the experiment with 100% of the tags and $\alpha = 0.8$ (80% of textual information).

Table 1. MAP of SINAI submitted experiments

α	20%	30%	40%	50%	60%	100%
0.1	0.0232	0.0265	0.0267	0.0262	0.0259	0.0260
0.2	0.0341	0.0414	0.0418	0.0399	0.0390	0.0410
0.3	0.0499	0.0620	0.0622	0.0603	0.0588	0.0638
0.4	0.0846	0.1019	0.1021	0.1000	0.0971	0.1067
0.5	0.1727	0.1899	0.1923	0.1886	0.1875	0.2075
0.6	0.2203	0.2567	0.2625	0.2587	0.2529	0.2700
0.7	0.2350	0.2734	0.2847	0.2785	0.2727	0.2889
0.8	0.2396	0.2768	0.2911	0.2885	0.2834	0.2996
0.9	0.2394	0.2749	0.2891	0.2879	0.2833	0.2988
1.0	0.2356	0.2710	0.2838	0.2828	0.2791	0.2944

Table 2 shows the best results for the groups participating in ImageCLEFmed 2007, only in the mixed retrieval subtask. Our best experiment is in the second position in the list. The OHSU group obtained the best results in mixed retrieval task. In their experiments [6] they used supervised machine learning techniques using visual features to classify the images based on image acquisition modality, in addition to the textual retrieval. Moreover, they analyzed the query with a Bayesian classifier to discern the desired image modality. The RWTH group used feature weights that were trained using the maximum entropy method with the topics of 2005 and 2006 jointly respectively [7].

In Table 1 we can see the results of only textual experiments in the $\alpha=0.1$ row. The best result obtained is 0.2944 of MAP value, using 100% of tags in the collection. The best results for the groups participating in only textual retrieval can be seen in Table 3. Our best experiment is in the fourth position in the list. The MRIM-LIG group took concepts from a textual corpus and queries, and they created a graph model with these concepts [8]. Therefore, they used a semantic model for the search. This model is not comparable with our system. The MIRACLE group only filtered the query and the collection with stemmer and stopper processes [9]. Finally, the OHSU group obtained their best result using the same approach as in mixed retrieval [6].

Table 2. Performance of official runs in mixed Medical Image Retrieval

Run	MAP
ohsu_m2_rev1.c.txt	0.3415
SinaiC100T80.clef	0.2999
RWTH-FIRE-ME-tr0506	0.2962
UB-NLM-UBTL3	0.2938
miracleMixGENTRIGHTmin.txt	0.2480
GE_VT1_4.treval	0.2195
CINDI_TXT_IMAGE_LINEAR.txt	0.1906
7fire-3easyir.clef	0.0224

Table 3. Performance of official runs in only text Medical Image Retrieval

Run	MAP
LIG-MRIM-LIG_MU_A	0.3538
miracleTxtENN.txt	0.3385
ohsu_text_e4_out_rev1.txt	0.3317
SinaiC100T100.clef	0.2944
UB-NLM-UBTextBL1	0.2897
IPAL-IPAL1_TXT_BAY_ISA0.2	0.2784
GE_EN.treval	0.2369
DEU_CS-DEU_R1	0.1611
IRIT_RunMed1	0.0486

6 Conclusions and Further Work

This year, two new collections have been included in the ImageCLEFmed 2007. Adding this new data in the collection improves the results obtained with the Lemur IR system compared to those from previous years. We believe that this is due to a wider set of samples provided by the large new sub-collection MyPACS.

However, the results of our experiments with expanded queries are not very good, since the expanded queries have lots of repeated words. These words do not add new information, but their weight increases in the query, which may not be desirable because the standard Lemur query parser assumes that a very frequent word is a very important word in the query.

In the mixed retrieval experiments, the system with similar or better results than ours (by the OHSU team) relies on a supervised learning phase. Our system does not need any kind of training. Nevertheless, the idea of classifying by image acquisition modality used in the OHSU experiments [6] is very interesting.

Our next step will focus on improving the query expansion and filtering the number of words included in the expansion. Moreover, we are considering the use of the UMLS ontology⁴ in order to include the multilingual features of

⁴ <http://www.nlm.nih.gov/pubs/factsheets/umls.html>

the collections. To improve the query expansion, we will use the extended query language of Lemur, giving a different weight to the expanded terms.

Furthermore, we will study a way to eliminate the images that are not in the same image acquisition modality expressed in the query, as does the OHSU team, but considering just textual information.

Acknowledgements

This project has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03).

References

- [1] Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersch, W.: Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. In: CLEF Workshop (2007)
- [2] 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, pp. 652–661. Springer, Heidelberg (2006)
- [3] Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T., Montejoraez, A., Ureña-López, L.A.: SINAI at ImageCLEF 2006. In: CLEF Workshop (2006)
- [4] Ogilvie, P., Callan, J.P.: Experiments Using the Lemur Toolkit. TREC (2001)
- [5] Chevallet, J.P., Lim, J.H., Radhouani, S.: A Structured Visual Learning Approach Mixed with Ontology Dimensions for Medical Queries. 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. 642–651. Springer, Heidelberg (2006)
- [6] Kalpathy-Cramer, J., Hersch, W.: Medical Image Retrieval and Automatic Annotation: OHSU at ImageCLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 623–630. Springer, Heidelberg (2008)
- [7] Deselaers, T., Gass, T., Weyand, T., Ney, H.: FIRE in ImageCLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 492–499. Springer, Heidelberg (2008)
- [8] Maisonnasse, L., Gaussier, E., Chevallet, J.P.: Multiplying Concept Sources for Graph Modeling. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 585–592. Springer, Heidelberg (2008)
- [9] Villena-Román, J., Lana-Serrano, S., González-Cristóbal, J.C.: MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 593–596. Springer, Heidelberg (2008)

Speeding Up IDM without Degradation of Retrieval Quality

Michael Springmann and Heiko Schuldt

Database and Information Systems Group, University of Basel, Switzerland
{michael.springmann,heiko.schuldt}@unibas.ch

Abstract. The Image Distortion Model (IDM) has previously shown good retrieval quality. However, one of the limitations that may limit its use in a wider range of applications is computational complexity. In this paper, we present an approach that applies several optimizations to decrease the retrieval time of IDM without degrading the quality of query results. We were able to perform the IDM in less than 1.5 seconds per query on an 8-way server and 16 seconds on a standard Pentium 4. In particular, the early termination strategy we applied contributed a speedup of up to 4.9. We also extended the possible displacements to an area of 7×7 pixels with a local context of up to the same size. The results submitted to the medical automatic annotation task of ImageCLEF'2007 were ranked in the upper third. Most importantly, the proposed techniques are not limited to IDM but can also be applied to other expensive distance measures.

1 Introduction

The Image Distortion Model (IDM) has shown very good retrieval quality in the Medical Automatic Annotation Task at ImageCLEF 2005 [1] and was still ranked in the top 10 results of ImageCLEF 2006 [2]. IDM performs a pixel-by-pixel value comparison of downscaled versions of the image on either gray values or gradients or a combination of both. It allows some displacements of pixels within the so-called warp range. Instead of plain pixels, IDM can also use a local context, e.g. a 3×3 area around a center pixel for comparison – which improves the query results, but is at the same time associated with significantly higher complexity. This complexity is even further increased when the warp range and local context are enlarged which, on the other hand, showed significant improvements in the quality of the retrieval results and of the subsequent classification of the medical images.

According to [3], the execution of a single query in a collection of 8'728 images with a local context of 3×3 pixels and a warp range of 5×5 pixels took about 5 minutes on a standard Pentium PC with 2.4 GHz. This is clearly not even close to interactive response time as it would be required for many applications in the area of content-based image retrieval and also too slow for image classification. The authors therefore proposed to use a sieve function to reduce the number of expensive IDM computations to a smaller subset of images, which have been pre-selected using the less expensive Euclidean distance. By this, they reduced

the time for a single query to 18.7 seconds, but this is only possible with some degradation in retrieval quality. Using such an approach therefore requires to find a good tradeoff between speed and quality.

We propose an approach that increases the speed of IDM without any negative impact on the retrieval quality. In our experiments, we used the parameters for IDM as in [3,4], but instead of the sieve function we apply an early termination condition in the individual distance computation. By this, we can reduce the execution time on similar hardware to 16 seconds per query without any degradation of quality. On modern hardware with 8 CPUs we could reduce this time further to less than 1.5 seconds per query by exploiting multithreading.

The main improvement in our approach is based on the following idea: As a result of the query, we want to retrieve a set of the k nearest neighbor images of the query image within a potentially large collection. This set is passed to a kNN classifier to determine the class of the query image. Images which are not among the k nearest neighbors do not contribute to the classification and their exact distance is not required. Therefore we abort the distance computation as soon as we can determine safely, that the distance will not be sufficiently small to become relevant for the k nearest neighbors. Such techniques are not limited to the image distortion model and have already been used in the VA-file [5], which is a part of the ISIS [6] content-based image similarity search system.

For our runs in the ImageCLEF Medical Automatic Annotation Task 2007, we extended the warp range of IDM to three pixels instead of only two and used a local context of 5×5 and 7×7 pixels instead of only 3×3 as in [4], which improved the retrieval score about 9.7%. Notice that extending the area results in significantly longer computation times. Therefore, the above mentioned optimizations are essential. Due to the fact that our approach does not need to degrade the performance by a sieve function and by using a modified classifier that better exploits the hierarchical structure of the IRMA code [7], we were able to reduce the error score by 12.6%.

The remainder of this paper is organized as follows: Section 2 introduces in depth the used distance measure and the chosen parameters. Section 3 describes the algorithmic optimizations, that achieved most of the performance improvements in the similarity computation. The results of all these modifications are presented in Section 4. Section 5 concludes.

2 The IDM Distance Measure and Parameters of the Submitted Runs

2.1 IDM with Local Context

The IDM is a deformation model [4] which is used as a measure of distance or dissimilarity between two images. By using a kNN classifier, it can be applied to the automatic medical image annotation task of ImageCLEF [7].

To determine the k nearest neighbors of a query image, the latter must be compared to each image of the collection using a distance measure. If we assume for

simplicity that every image is scaled down to 32 pixels width and height, we can interpret its gray values as a feature vector of length 1024. The Euclidean distance is an example for an elementary distance measure that can be applied to such images.

IDM allows for some displacements of each individual pixel on both axes within the range of the warping distance w . Each displacement may get penalized with some costs that are associated with this displacement which are computed using the cost function C . If we consider only the possible displacements of a single pixel within the warp range w , always the one is chosen that results in the smallest distance. The warp range is illustrated as the inner area in Fig. 1 on the reference image R.¹

We use a threshold t to limit the maximum contribution of a single pixel as in [8] which improved retrieval quality on the test data set. The retrieval quality of IDM is significantly increased, if IDM is not limited to single pixels, but their local context [9] is considered as well. The local context is defined by an area of pixels around the central pixel that differ in their row and column value by not more than the local context range l . IDM with local context computes the average distance between those pixels in the area with the corresponding pixels of the reference image.

Due to the local context, we preserve the aspect ratio of images when scaling them: We use images where the longer side has no more than 32 pixels, which is still sufficient for finding nearest neighbors for classifying the image. In order to make images of different width or height comparable, we identify each corresponding pixel by scaling. The local context is illustrated as the area in Fig. 1 on the query image Q. For the corresponding pixel in the reference image R, the local context surrounds the top left pixel within the warp range and will be moved according to the small arrows to minimize the distance within the warp range.

The local context is always taken around the corresponding pixels directly from the image without adjusting further the dimensions of the query and reference image, such that the directionality of edges is preserved in the local context of pixels. Such edges can be detected using a Sobel filter. For the computation of IDM, either the gray values of the pixels can be used directly as shown in Fig. 1, or their gradients (detected edges), or a combination of both.

2.2 Modifications of Parameters

The recommended parameters in [4] are: $w \in \{1, 2\}$, allowing a deformation in a warp range of 3×3 or 5×5 pixels and a local context range $l = 1$, hence using another 3×3 pixel area as local context. For our runs, we used $w = 3$, therefore allowing displacements in an area of 7×7 pixels. The area covered by the original parameters is illustrated in Fig. 1 on the left part of the reference image R, the area covered by the extended parameters is illustrated on the right part.

We increased also the local context range to $l = 2$ for most runs, our best run used $l = 3$. We applied a cost function with two different sets of parameters, out of which

¹ There is no corresponding area in the query image Q on the right-hand side, since for single pixel comparison, one pixel is mapped to the best-fitting corresponding pixel of R within the warp range.

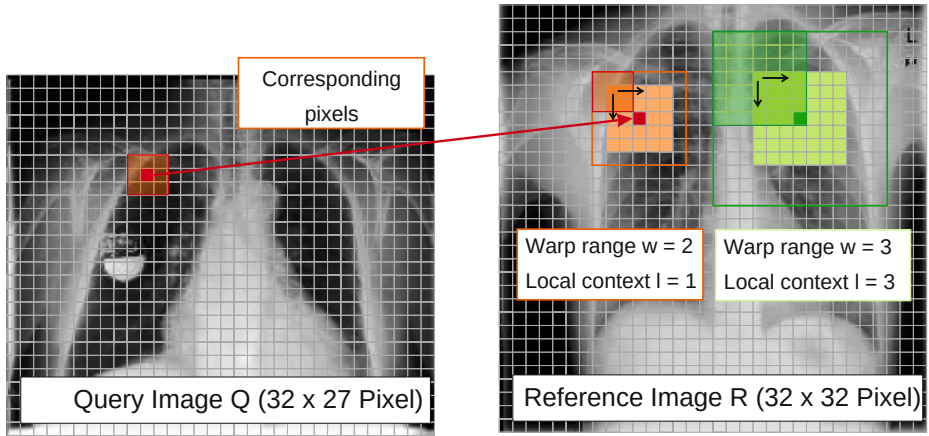


Fig. 1. Example of areas affected by comparison of pixel with IDM

the one with higher costs and therefore higher penalty for displacements achieved slightly better results. In contrast to [10], we use both layers: The gray values of down scaled images directly (intensities) and also the Sobel-filtered image (gradients), where the gray values have twice the importance of the Sobel-filtered version.

Finally, we used a kNN classifier with $k \in \{3, 5\}$ that not only takes the class of the nearest neighbors into account, but also weighs the class information based on the inverse square of the distance [11], which was computed using IDM. We further experimented with a modification that uses more knowledge about the IRMA code [7]: In case the distance is very close, but the classes of nearest neighbors do not match, the IRMA code including the ‘*’-symbol for each non-matching character is generated. Thus, uncertainty is expressed for cases where the closest neighbors differ. According to our experiments, slight improvements are possible, but the choice of thresholds is crucial. Since the classification does not affect the time needed to identify the k nearest neighbors, it will not be further discussed in this paper.

The baseline run [10] for the automatic medical image annotation task of ImageCLEF 2007 uses IDM in combination with Tamura texture features and the cross correlation-function (CCF) as it has been used the last two years. IDM was given the biggest weight in the baseline run and has been identified as the bottleneck for retrieval speed [3]. We focused only on IDM, leaving aside possible combinations with other features and distance measures.

3 Algorithmic Optimization

3.1 Early Termination Strategy

The distance of two images is aggregated out of the individual distances of the corresponding pixels in both images. This is independent of the distance measure used, i.e., IDM or Euclidean distance. This aggregation is performed

by processing one pixel at a time, thus this incrementally sums the full distance. Since the distance for a single pixel can never be negative, the aggregated sum will monotonically increase with each subsequent pixel.

As the result of the nearest neighbor search, only the ordering and distance of the k nearest neighbors will be used in the classification step. Therefore the exact distance of any image with a rank $> k$ is unimportant and can safely be approximated or discarded. We keep an ordered list based on the distances that have been evaluated so far. This list is restricted in size to k and is updated by inserting each newly computed distance. At the same time, items with a rank that would exceed the size of the list are dropped. Every image that is part of the final k nearest neighbors must achieve a distance less or equal to the last entry of this list. Therefore we derive out of this value the maximum sum as a threshold for computation of the distance of the next image.

After each pixel of the query image has been processed, the current sum is compared to the maximum sum. If it exceeds the maximum, computation is terminated for this image. If not, computation is continued until all pixels have been processed and the distance and a reference to the image are inserted into the list of distances at the appropriate position. Any item above the k^{th} position is dropped from the list. After all reference images have been processed, this list contains the result of the nearest neighbor search.

Since the algorithm is orthogonal to the used function to compute the distance of a single pixel in the query image, it can be applied to IDM as well as the Euclidean distance or any similar distance function.

3.2 Multithreading

Within the last years, multi-core CPUs became very popular and affordable. Therefore it becomes more and more important to design applications in a way that they can use multiple threads in order to utilize all the capabilities provided by current hardware.

In our implementation, a dispatcher takes the computed result, updates the list of found distances as described in Section 3.1 and assigns the next task to the thread. By this we could achieve almost linear speedup on multi-core CPUs. Our approach is similar to the “parallelization on second level” described in [12], except that we did not use OpenMP but plain Java threads and that our dispatcher enforces sequential reads from disk and provides all required support for the early termination strategy in concurrent computation.

4 Results

For the submitted runs, we performed all experiments on IBM xSeries 445 with 8 Xeon MP CPUs. We used an image distortion model with a warp range of $w = 3$ and two different cost matrices, 4812 and 369, where the one assigning higher costs (4812) performed slightly better when using only a 5×5 pixel area ($l = 2$) as local context. For $l = 3$ we submitted only a single run using the 369 matrix

Table 1. Scores and execution times of performed runs on 8-way Xeon MP 2.8 GHz for entire run with 1.000 queries

Rank	Run id - UNIBAS-DBIS-	w	l	k	Score	Execution time
19	IDM_HMM_W3_H3_C	3	3	3c	58.15	3h 41m
N/A	IDM_HMM_W3_H3	3	3	3	59.12	3h 41m
20	IDM_HMM2_4812_K3	3	2	3	59.84	2h 38m
21	IDM_HMM2_4812_K3_C	3	2	3c	60.67	2h 38m
22	IDM_HMM2_4812_K5_C	3	2	5c	61.41	2h 38m
23	IDM_HMM_369_K3_C	3	2	3c	62.83	2h 23m
24	IDM_HMM_369_K3	3	2	3	63.44	2h 23m
25	IDM_HMM_369_K5_C	3	2	5c	65.09	2h 23m
N/A	IDM	2	1	5	65.45	30m 12s
N/A	IDM	2	1	1	66.17	23m 2s
N/A	+Sieve Euclid $c = 500$	2	1	1	66.50	6m 39s

– which turned out to be the best of our runs. A “c” appended to the number of k nearest neighbors indicates that the IRMA-code aware classifier was used. We used a per-pixel threshold of $t = 96$ in all cases. “N/A” indicates that no official rank was assigned since this run was not submitted to the ImageCLEF benchmark.

The achieved scores presented in Table 1 show that:

- Increasing the warp range and local context improves retrieval quality.
- Classification with $k = 3$ outperforms $k = 5$. Further experiments showed that $k = 1$ is inferior to both when the inverse square of the distance is used. Expressing uncertainty in the IRMA-code pays off only when a threshold on the distance is set very carefully. In our best case, this improved the score by 1.6% with only the two nearest neighbors contributing to the generated code if below the threshold, otherwise the 3 nearest neighbors were used for the traditional kNN classifier.
- Using the sieve function proposed in [3] slightly degrades retrieval quality.
- Overall the score was improved by 12.56% from 66.50 to 58.15.

The execution times have been measured for entire runs on 1’000 images, features were cached in main memory. The average execution time for interactive queries can simply be derived by dividing the time of the entire run by 1’000, since we did not apply additional optimizations e.g. extract features of the next query image while the last one is still being processed.

This means for $l = 2$, a single query took on average less than 8.6 seconds and 13.3 seconds for $l = 3$. For comparison: Our best run with $w = 3$ and $l = 3$ took 16h 18m when no early termination based on the maximum sum was used. This long duration even increased to entire 5 days, 17h and 32m on the same machine when we limited the number of used threads to a single one – as it was the case in our starting point of the implementation. This means that our optimizations achieved a speedup of 4.42 and 37.34, respectively.

We also performed runs with the parameters proposed in [4]: IDM with a deformation in 5×5 pixels ($w = 2$) and a local context of 3×3 pixels ($l = 1$) and

the nearest neighbor decision rule ($k = 1$). On a standard Pentium 4 PC with 2.4 GHz, this run finished within 4 hours and 28 minutes (16.0 seconds per query) – without any sieve function. The same run was finished on the 8-way Xeon within 23 minutes and 2 seconds (less than 1.5 seconds per query). When we turned off just the early termination, the durations increased to 1 hour 59 minutes on an 8-way Xeon MP server (7.1 seconds per query, factor 4.86) and 19 hours 21 minutes on a Pentium 4 (69.77 seconds per query, factor 4.33). When features were read from disk directly and not cached between several queries, the IDM computation using the sieve function degraded significantly to 41 minutes and 50 seconds (2.5 seconds per query) and therefore did not perform much faster than the plain IDM computation in the same setting, which took 46 minutes and 53 seconds (2.8 seconds per query).

5 Conclusion and Outlook

Increasing the warp range from 2 to 3 and using a local context to an area of 7×7 pixels instead of 3×3 significantly improves the retrieval quality of IDM. Modifications of the used kNN classifier can further improve the quality, but in all our experiments only to a much smaller extent.

For being able to perform such experiments within reasonable time, we propose an early termination strategy, which has proven to successfully speed up the expensive Image Distortion Model by factors in the range of 4.3 to 4.9 in our experiments. We could reduce the computation time to perform a single query in the collection of 10'000 reference images to approximately 16 seconds on a standard Pentium 4 PC. Making proper use of multithreading allows to perform a single query within 1.5 seconds on an 8-way Xeon MP server. Even with the increased range, we could perform this on the Xeon server in maximum 13.3 seconds per query.

Acknowledgments

Part of the implementation are based on the work performed by Andreas Dander at the University for Health Sciences, Medical Informatics and Technology (UMIT), Hall i.T., Austria.

References

1. Deselaers, T., Müller, H., Clough, P., Ney, H., Lehmann, T.M.: The CLEF 2005 automatic medical image annotation task. *International Journal of Computer Vision* 74(1), 51–58 (2007)
2. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)

3. Thies, C., Güld, M.O., Fischer, B., Lehmann, T.M.: Content-based queries on the casimage database within the IRMA Framework: A field report. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 781–792. Springer, Heidelberg (2005)
4. Keysers, D., Deselaers, T., Gollan, C., Ney, H.: Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(8), 1422–1435 (2007)
5. Weber, R., Schek, H.J., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: VLDB 1998, Proceedings of 24th International Conference on Very Large Data Bases, pp. 194–205. Morgan Kaufmann, San Francisco (1998)
6. Brettlecker, G., Ranaldi, P., Schek, H.J., Schuldt, H., Springmann, M.: ISIS & OSIRIS: a process-based digital library application on top of a distributed process support middleware. In: DELOS Conference 2007 Working Notes, February 2007, pp. 81–90 (2007)
7. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Hersh, W.: Overview of the ImageCLEF 2007 medical retrieval and annotation tasks. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 472–491. Springer, Heidelberg (2008)
8. Keysers, D., Dahmen, J., Ney, H., Wein, B.B., Lehmann, T.M.: Statistical framework for model-based image retrieval in medical applications. *Journal of Electronic Imaging* 12, 59–68 (2003)
9. Keysers, D., Gollan, C., Ney, H.: Local context in non-linear deformation models for handwritten character recognition. In: ICPR 2004: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR 2004), Washington, DC, USA, vol. 4, pp. 511–514. IEEE Computer Society, Los Alamitos (2004)
10. Güld, M.O., Deserno, T.M.: Baseline results for the CLEF 2007 medical automatic annotation task. In: Working Notes of the CLEF Workshop 2007 (September 2007)
11. Esters, M., Sander, J.: *Knowledge Discovery in Databases*. Springer, Berlin (2000)
12. Terboven, C., Deselaers, T., Bischof, C., Ney, H.: Shared Memory Parallelization for Content-based Image Retrieval. In: ECCV 2006 Workshop on Computation Intensive Methods for Computer Vision, Graz, Austria (May 2006)

Content-Based Medical Image Retrieval Using Low-Level Visual Features and Modality Identification

Juan C. Caicedo, Fabio A. Gonzalez, and Eduardo Romero

BioIngenium Research Group
National University of Colombia
{jccaicedoru,fagonzalezo,edromero}@unal.edu.co
<http://www.bioingenium.unal.edu.co>

Abstract. This paper presents the image retrieval results obtained by the BioIngenium Research Group, in the frame of the ImageCLEFmed 2007 edition. The applied approach consists of two main phases: a pre-processing phase, which builds an image category index and a retrieval phase, which ranks similar images. Both phases are based only on visual information. The experiments show a consistent frame with theory in content-based image retrieval: filtering images with a conceptual index outperforms only-ranking-based strategies; combining features is better than using individual features; and low-level features are not enough to model image semantics.

1 Introduction

Designing and modeling methods for medical image search is a challenging task. Hospitals and health centers are surrounded by a large number of medical images with different types of contents, which are mainly archived in traditional information systems. In the last decade, content-based image retrieval methods have been widely studied in different application domains [1] and particularly, research in the medical field has taken special interest. The ImageCLEFmed is a retrieval challenge in a collection of medical images [2], which is organized yearly to stimulate the development of new retrieval models for heterogeneous document collections containing medical images as well as text. The BioIngenium Research Group at the National University of Colombia participated in the retrieval task of the ImageCLEFmed 2007 edition [3], using only visual information.

Some important issues for retrieving in heterogeneous image collections are a coherent image modeling and a proper problem understanding. Different modalities of medical images (radiography, ultrasound, tomography, etc.) could be discriminated using basic low level characteristics such as particular colors, textures or shapes as they are at the base of most image analysis methods. Traditional approaches are mainly based on low-level features which describe the visual appearance of images, because those descriptors are general enough to

represent heterogeneous contents [4]. Histogram features and global descriptors have been used to build similarity measures between medical images, obtaining poor results in heterogeneous collections because they do not fully describe the content's semantic.

This work attempts to introduce some slight additional information in the retrieval process by the use of a filtering method. Our approach, firstly try to identify a general modality for each image in the database in a pre-processing phase and then uses histogram features for ranking. This two-phase approach makes use of low-level features to describe image contents and a classification model to recognize the modality associated to one image. We accepted the sytem should recognize 11 general image modalities so that the retrieval algorithm was forced to a subset of images conceptually related to the query. In this paper some details about the system and the model used by the BioIngenium Research Group to participate in the ImageCLEFmed 2007 edition are presented and discussed. The reminder of this paper is organized as follows: Section 2 presents the two-phase proposed approach. Section 3 presents and discusses the results obtained in the challenge evaluation and Section 4 contains some concluding remarks and future work.

2 Proposed Approach

The image retrieval process consists of two main phases: pre-processing phase and retrieval phase. Both phases are described as follows.

2.1 Pre-processing Phase

The pre-processing phase is composed of two main components: a feature extraction model and a classification model. The input of the pre-processing phase is the original image database, i.e. images from the ImageCLEFmed collection, with more than 66,000 medical images. The output of the pre-processing phase is an index relating each image to its modality and a feature database. This scheme is shown in Figure 1.

The Feature Extraction Model. The feature extraction model operates on the image database to produce two kind of features: histogram features and meta-features. Histogram features are used to build the feature database, which is used in the retrieval phase to rank similar images. Meta-features are a set of histogram descriptors, which are used as the input to the classification model to be described later. Histogram features used in this system are [4,5,6]:

- *Gray scale and color histogram (Gray and RGB)*
- *Local Binary Partition histogram (LBP)*
- *Tamura texture histogram (Tamura)*
- *Sobel histogram (Sobel)*
- *Invariant feature histogram (Invariant)*

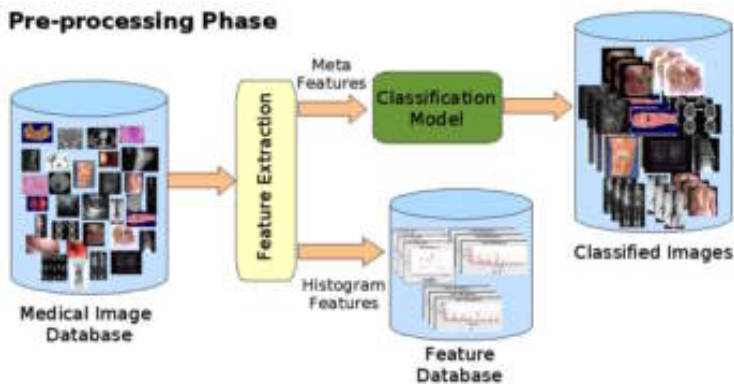


Fig. 1. Preprocessing phase: The input corresponds to a medical image database. The phase produces as output the feature database and the classified images. This phase uses a low-level feature extraction framework and a classification model based on a multilayer perceptron or a support vector machine.

Metafeatures are calculated from histogram features in order to reduce the dimensionality. These metafeatures are the four moments of the moment generating function (mean, deviation, skewness and kurtosis) and the entropy of the histogram. Each histogram has five associated metafeatures, meaning a total of 30 meta-features with information of color, texture, edges and invariants.

Classification Model. Since the data set contains different type of images with different modalities, the proposed approach first attempts to identify the modality of a given image. This restricts the query results to contain images with the same modality as the query image. The classifier is not applied to the raw information of the histograms, since the dimensionality of the feature vector will be very high. Instead, the set of metafeatures are used to reduce the dimensionality, with some information loss. A machine-learning approach is used to classify images in the database. First a training set was selected from the database composed of 2,500 images in 11 categories, each category corresponding to a general image modality . Image modalities are described in Table 1.

This dataset was used as training set for two classifiers. The first classifier is a Support Vector Machine (SVM) with the Gaussian kernel [7]. The second classifier is a Multilayer Perceptron (MP) with one hidden layer and a variable number of

Table 1. Image categories

Category	Examples	Category	Examples	Category	Examples
Angiography	98	Histology	401	Magnetic Resonance	382
Ultrasound	183	Organ photo	196	Tomography	364
Endoscopy	137	Patient photo	171	Drawing	117
Gamagraphy	159	Radiography	344		

neurons, 30 inputs and 11 outputs. Each classifier had a training phase in which the hyper-parameters (complexity for the SVM and number of hidden neurons for the MP) were tuned, using 10-fold cross validation. A test set of images was used to calculate an estimate of the classification error on unseen instances.

Table 2 shows the performance of the best classification models in both training and test sets.

Table 2. Performance of the modality classification models on training and test sets

	Parameters	Training set error	Test set error
Multilayer Perceptron	Hidden nodes: 40	11.83%	20.78%
Support Vector Machine	$\gamma = 1, \lambda = 2$	18.69%	22.10%

2.2 Retrieval Phase

The image ranking process starts by receiving the query. The first step is to classify this image in order to restrict the search only to images with the same modality, herein called the filtering method. Then, the relevance ranking is calculated using different similarity measures.

Filtering. Images in the database are filtered according to the modality of the query image. For this purpose, the query image is classified, using the model trained in the pre-processing phase.

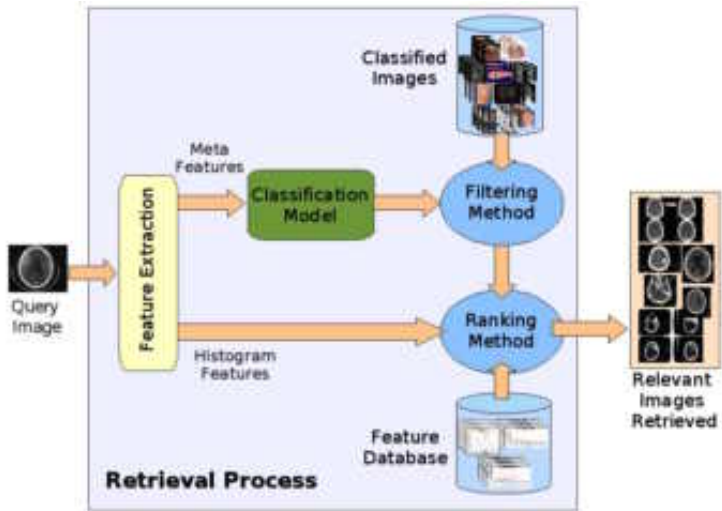


Fig. 2. Retrieval phase: A query image is received as input and a set of relevant images is generated as output. This phase uses the same feature extraction framework as in the pre-processing phase, but only applied to the query image. It also uses the previously trained classification model on the query image to select the subset of images to rank.

Ranking. Images are represented in this phase as histograms so that distances are calculated using similarity measures. In this work, five different similarity measures were tested: Euclidean distance, Relative Bin Deviation, Relative Deviation, Chi-square distance and Jhensen-Shannon Divergence; the last four specifically designed for histogram comparison.

The scheme has been adapted from our previous work on content-based image retrieval in a histopathology-image domain [8,9]. In that work, all the combinations of histogram types and similarity measures were tested to choose the best-performing similarity measure for each type of histogram.

Specifically, for each feature-similarity pair the retrieval performance on a set of images was calculated. The better feature-similarity combinations for each histogram feature are shown in Table 3. The similarity measures that produced the best results were Jensen-Shannon Divergence and Relative Bin Deviation. These similarities are defined as follows:

Jensen-Shannon Divergence

$$D_{JSD}(H, H') = \sum_{m=1}^M H_m \log \frac{2H_m}{H_m + H'_m} + H'_m \log \frac{2H'_m}{H'_m + H_m} \quad (1)$$

Relative Bin Deviation

$$D_{rbd}(H, H') = \sum_{m=1}^M \frac{\sqrt{(H_m - H'_m)^2}}{\frac{1}{2}(\sqrt{H_m} + \sqrt{H'_m})} \quad (2)$$

where M is the number of bins in the histogram and H_m is the value of the m -th bin.

Table 3. Feature-metric pairs defined as similarity measures on image contents

Feature-similarity	Feature-similarity
Gray-RelativeBinDeviation	LBP-RelativeBinDeviation
RGB-RelativeBinDeviaton	Tamura-RelativeBinDeviation
Sobel-JhensenShannon	Invariant-JhensenShannon

Similarity Measure Combination. The combination of multiple similarity measures may produce better results than using the individual measures. To combine similarity measures we used a Cross Category Feature Importance (CCFI) scheme [10]. This scheme uses the probability distribution of metafeatures to calculate a weight for each similarity measure. The combined similarity measure is:

$$s(x, y) = \sum_{f \in F} \omega(f) s_f(x, y) \quad (3)$$

where x and y are images, F is the feature set, $s_f(\cdot)$ is the similarity measure associated to the feature f and $\omega(f)$ is the importance factor for that similarity measure. The CCFI calculates each ω in the following way:

$$\omega(f) = \sum_{c_j \in J} p(c_j | f)^2 \quad (4)$$

Since we have some categories predefined in the database, we can calculate the weight of each feature using the probability class distribution of features. There are two classifications produced by different classifiers: SVM classification and MP classification. In each case the probability distribution varies according to the final classification. That means that the weights calculated in the scenario of the SVM classifier are different of those calculated in the scenario of the MP classifier.

3 Results and Discussion

3.1 Experimental Settings

We sent eight runs for evaluation that are divided into two groups: one using the MP classifier and the other using the SVM classifier. That is to say, the filtering method in the retrieval phase depends on the selected classifier. As each group of experiments have four runs, they correspond to four different strategies in the ranking method. Although our system have six similarity measures implemented, we sent three runs using only three of them individually: RGBHisto-RBD, Tamura-RBD, Sobel-JS. The fourth run corresponds to the similarity measure combination, that operates with the six implemented similarity measures.

3.2 Results

The results of our eight experiments are shown in Table 4, sorted out by MAP. In this table, the column *Run* shows the name of the sent experiment, following a three-parts convention: (1) *UNALCO* to identify our group at the National University of Colombia; (2) an identifier for the classification model used, *nmi* for the multilayer perceptron and a *svmRBF* for the support vector machine; and (3) the name of the filtering method used: RGB histogram (RGBHisto), Sobel histogram (Sobel), Tamura histogram (Tamura), and lineal combination of features (FeatComb).

The general ranking of our runs follows what is currently considered as true. Firstly, the MP classifier used for image filtering together with a feature-combination strategy for image ranking, shows the best MAP score in this set of runs. In all cases, the MP classifier shows better performance than the SVM to filter images, which is in general consistent with the error rates obtained in the training phase (Table 2). Tamura texture shows the worst results in both filtering strategies. In general, the feature combination approach performs better than individual similarity measures, suggesting that the combination strategy using the *Cross Category Feature Importance* scheme is a useful approach that effectively combine features based on their probability distribution.

Table 4. Automatic runs using only visual information

Run	Relevant	MAP	R-prec	P10	P30	P100
UNALCO-nni_FeatComb	644	0.0082	0.0149	0.020	0.0144	0.0143
UNALCO-nni_RGBHisto	530	0.0080	0.0186	0.0267	0.0156	0.0153
UNALCO-nni_Sobel	505	0.0079	0.0184	0.020	0.0167	0.0187
UNALCO-nni_Tamura	558	0.0069	0.0167	0.0233	0.0156	0.0153
UNALCO-svmRBF_Sobel	344	0.0056	0.0138	0.0033	0.0133	0.0133
UNALCO-svmRBF_FeatComb	422	0.0051	0.0077	0.010	0.0089	0.0093
UNALCO-svmRBF_RGBHisto	368	0.0050	0.0103	0.0133	0.010	0.0093
UNALCO-svmRBF_Tamura	375	0.0048	0.0109	0.0067	0.010	0.010

3.3 Discussion

The performance of the proposed approach in the competition is actually not enough for medical image retrieval. This could be explained, in particular by the fact that a restricted set of features was used¹ and, in general, by the fact that visual features alone are not enough for achieving a good retrieval performance.

In general, results behave as we expected: low-level features are still poor to describe the medical image semantics. Nevertheless, those results show that our scheme is consistent with general concepts in content-based image retrieval. First, the feature combination strategy performs better than the individual feature approach, suggesting that visual concepts can be modeled by mixing low-level features. Second, the filtering strategy allows a better retrieval than a simple one i.e. only-visual approaches (GE_GIFT and DEU_CS groups). Furthermore, a good filtering strategy allows identification of more relevant images. In fact, the SVM classification model performs poorer than the MP classification model in the training and testing sets and this could be related to the worst retrieval performance of the SVM-based runs.

4 Conclusions and Future Work

Content-based medical image retrieval is still a challenging task that needs new and clever methods to implement useful and effective systems. This paper discusses the main components of an image-retrieval system based on a two-phase strategy to build an image category index and to rank relevant images. This system is completely based on low-level visual information and makes not use of textual data. In general, obtained results match well with what one would expect, not only because of the well known semantic gap but because of the consistency in feature combination and filtering quality.

The future work at our lab will aim to take full advantage of all information into the collection, i.e. to involve textual data. Although textual data alone

¹ Content-based image retrieval systems such as GIFT and FIRE use considerably more visual features than our approach.

has demonstrated to be successful for image retrieval, we are very interested in models that mix up textual and visual data to improve the performance of our retrieval system.

References

1. Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. Technical report, Intelligent Sensory Information Systems, University of Amsterdam (2000)
2. Müller, H., Michoux, N., Bandon, D., Geissbühler, A.: A review of content based image retrieval systems in medical applications clinical benefits and future directions. *International Journal of Medical Informatics* 73, 1–23 (2004)
3. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Hersh, W.: Overview of the imageclef 2007 medical retrieval and annotation tasks. *Cross-Language Retrieval in Image Collections (ImageCLEF)* (2007)
4. Deselaers, T.: Features for Image Retrieval. PhD thesis, RWTH Aachen University. Aachen, Germany (2003)
5. Siggelkow, S.: Feature Histograms for Content-Based Image Retrieval. PhD thesis, Albert-Ludwigs-Universität Freiburg im Breisgau (2002)
6. Mark, S., Nikson, A.S.A.: Feature Extraction and Image Processing. Elsevier, Amsterdam (2002)
7. Schölkopf, B., Smola, A.: Learning with kernels. Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2002)
8. Caicedo, J.C., Gonzalez, F.A., Romero, E., Triana, E.: Design of a medical image database with content-based retrieval capabilities. In: *Advances in Image and Video Technology. IEEE Pacific Rim Symposium on Image Video and Technology. PSIVT 2007* (2007)
9. Caicedo, J.C., Gonzalez, F.A., Romero, E., Triana, E.: A semantic content-based retrieval method for histopathology images. In: *Information Retrieval Technology: Theory, Systems and Applications. Proceedings of the Asia Information Retrieval Symposium, AIRS 2008* (2008)
10. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review* 11, 273–314 (1997)

Medical Image Retrieval and Automatic Annotation: OHSU at ImageCLEF 2007

Jayashree Kalpathy-Cramer and William Hersh

Department of Medical Informatics & Clinical Epidemiology
Oregon Health and Science University, Portland, OR, USA
{kalpathy, hersh}@ohsu.edu

Abstract. Oregon Health & Science University participated in the medical retrieval and medical annotation tasks of ImageCLEF 2007. In the medical retrieval task, we created a web-based retrieval system built on a full-text index of both image and case annotations. The text-based search engine was implemented in Ruby using Ferret, a port of Lucene and a custom query parser. In addition to this textual index of annotations, supervised machine learning techniques using visual features were used to classify the images based on image acquisition modality. All images were annotated with the purported modality. Purely textual runs as well as mixed runs using the purported modality were submitted, with the latter performing among the best of all participating research groups. In the automatic annotation task, we used the 'gist' technique to create the feature vectors. Using statistics derived from a set of multi-scale oriented filters, we created a 512-dimensional vector. PCA was then used to create a 100-dimensional vector. This feature vector was fed into a two layer neural network. Our error rate on the 1000 test images was 67.8 using the hierarchical error calculations.

1 Medical Image Retrieval

Advances in digital imaging technologies and the increasing prevalence of Picture Archival and Communication Systems (PACS) have led to a substantial growth in the number of digital images stored in hospitals and medical systems in recent years. In addition, on-line atlases of images have been created for many medical domains including dermatology, radiology and gastroenterology. Medical images can form an essential component of a patient's health record. Medical image retrieval systems can be important with aiding in diagnosis and treatment. They can also be highly effective in health care education, for students, instructors and patients.

1.1 Introduction

Image retrieval systems do not currently perform as well as their text counterparts [1]. Medical and other image retrieval systems have historically relied on annotations or captions associated with the images for indexing the retrieval system. The last few decades have seen numerous advancements in the area of content-based image retrieval (CBIR) [2,3]. Although CBIR systems have demonstrated success in fairly

constrained medical domains including pathology, dermatology, chest radiology, and mammography, they have demonstrated poor performance when applied to databases with a wide spectrum of imaging modalities, anatomies and pathologies [1,4,5,6].

Retrieval performance has shown demonstrable improvement by fusing the results of textual and visual techniques. This has especially been shown to improve early precision [7,8]. The medical image retrieval task within ImageCLEF (ImageCLEFmed) 2007 campaign is a TREC-style [9] and provides a forum and set of test collections for the medical image retrieval community to use to benchmark their algorithms on a set of queries. The ImageCLEF campaign has, since 2003, been a part of the Cross Language Evaluation Forum (CLEF) [9,10,11] which is derived from the Text Retrieval Conference (TREC, trec.nist.gov).

1.2 System Description of Our Adaptive Medical Image Retrieval System

The ImageCLEF collection consists of about 66,000 medical images and annotations associated with them. We have created a flexible database schema that allows us to easily incorporate new collections while facilitating retrieval using both text and visual techniques. The text annotations in the collection are currently indexed and we continue to add indexable fields for incorporating visual information.

Database and Web Application. We used the Ruby programming language, with the open source Ruby On Rails web application framework^{1, 2}. A PostgreSQL relational database was used to store the images and annotations.

The database has images from the four different collections that were part of the ImageCLEFmed 2006 image retrieval challenge as well as two new collections for 2007. The approximately 66,000 images in these collections reside in cases, with annotations in English, German and/or French. The collections themselves are substantially heterogeneous in their architectures. Some collections have only one image per case while others have many images per case. Annotation fields are also quite different among the collections. Some collections have case-based annotations while others have image-based annotations. This difference is especially significant for text based retrieval as images of different modalities or anatomies or pathologies could be linked to the same case annotation. In this situation, even though only one image from a case containing many images might be relevant to a query (based on the annotation), all images for the case would be retrieved in a purely text based system, reducing the precision of the search.

We used the relational database to maintain the mappings between the collections, the cases in the collections, the case-based annotations, the images associated with a collection, and the image based annotations.

Image Processing and Analysis. The image itself has important visual characteristics such as color and texture that can help in the retrieval process. Images that may have had information about the imaging modality or anatomy or view associated with them as part of the DICOM header can lose that information when the image is compressed

¹ <http://www.ruby-lang.org>

² <http://www.rubyonrails.org>

to become a part of a teaching or on-line collection, as the image format used by these collections is usually compressed JPEG.

We created additional tables in the database to store image information that was created using a variety of image processing techniques in MATLAB³. For instance, the images in the collection typically do not contain explicit details about the imaging modality. In previous work [8], we have described our modality classifier that can identify the imaging modality for medical images with a high level of confidence (>95% accuracy on the database used for the validation). Grey scale images are classified into a set of modalities including x-rays, CT, MRI, ultrasound and nuclear medicine. Color image classes include gross pathology, microscopy, and endoscopy.

Each image was annotated in the database with the purported image modality and a confidence value. This can be extremely useful for queries where the user has specified a desired image modality. An example query from ImageCLEF 2006 was “*Show me microscopic images of tissue from the cerebellum.*”

The precision of the result of such a query can be improved significantly by restricting the images returned to those of the modality desired [8]. This is especially useful in eliminating images of the incorrect modality that may be part of a case containing a relevant image from the returned list of images. However, this increase in precision may result in a loss in recall if the classification algorithm incorrectly classifies the image modality.

We continue to experiment with a variety of image clustering and classification algorithms and adding the numerical data and labels to the database. Clustering images that look visually similar can be again used to improve the precision of the image retrieval process and speed up the system searching of images in the same cluster as the query image (if available).

Query Parser and Search Engine . The system presents search options to the user including Boolean OR, AND and exact match. There are also options to perform fuzzy searches and custom query parsing. The cornerstone of our system is the query parser, written in Ruby. Ferret, a Ruby port of the popular Lucene system, was used in our system as the underlying search engine⁴.

Queries were first analyzed using MedPost, a Parts-of-Speech (POS) Tagger created using the Medline corpus, and distributed by the National Library of Medicine⁵ [14].

A simple Bayesian classifier⁶ was trained to discern the desired image modality from the query, if available. The classifier performed extremely well within the constrained vocabulary of imaging modalities. Stop words were then removed from the query. These include Standard English stop words as well as a small set of stop words determined by analyzing queries from the last three years, including ‘finding’, ‘showing’, ‘images’, ‘including’ and ‘containing’.

The system is also linked to the UMLS Metathesaurus. The user can choose to perform automatic query expansion using synonyms from the Metathesaurus.

³ <http://www.mathworks.com>

⁴ <http://ferret.davebalmmain.com>

⁵ <ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost>

⁶ <http://classifier.rubyforge.org>

A sample query “Show me CT images with a brain infarction” is automatically parsed and the following information is extracted from it: CT-> imaging modality, brain -> anatomic location, infarction -> finding. This information can be used to combine the results of the textual and visual systems more effectively.

1.3 Runs Submitted

We submitted a total of 10 runs. These runs included textual and mixed, automatic and manual options. The text runs had an “as-is” run, where the topics were submitted directly to the search system, a run where term expansion using the UMLS system was used, a text run where both our custom parser and query expansion was used and a manual run. We also submitted runs using different weighted combinations of the FIRE baseline (published by the organizers) with our baseline textual runs.



Fig. 1. Screen display of our system displaying user options

1.4 Results and Discussion

The complete performance of our runs can be found among the official ImageCLEFmed results. However, we note that there was a discrepancy between the order in the output of our image retrieval system and that which is required for trec_eval. In calculating the mean average precision (MAP), trec_eval only considers the “similarity score” column. Both the rank column and the order of the documents in the submission are ignored. Ties are broken lexicographically. Many participants, including OHSU, had created an ordered list of images, where the order in which the documents (images) appeared was considered the ranking of the documents. However, the score was either increasing from the top of the list to the bottom of the list or a number that was not unique or indicative of the desired ranking. This poor formatting of the submissions led to surprisingly poor

performance of some combined runs as well as runs of certain participants. This was discovered during the post workshop analysis of the results. The official runs, including the OHSU runs were reformatted where the score was set equal to 1/row order. This ensured that the score was in decreasing order from the top of the ordered list to the bottom of the ordered list, as required by trec_eval. Table 1 presents the official mean average precision (MAP) as well as the results of trec_eval on the reformatted runs for the most significant runs submitted by OHSU.

Table 1. Performance of significant OHSU runs

Run	Type	Official MAP	Reformatted MAP	Comments
ohsu_m2_rev1_c.txt	AM	0.341	0.408	mixed run using modality, starting from OHSU_txt_exp2
OHSU-oshu_man2	MT	0.346	0.360	manual run, using terms from umls expansion
ohsu_text_e4_out_rev1_c.txt	AT	0.332	0.347	query expansion and query parsing
OHSU-OHSU_txt_exp2	AT	0.319	0.334	query expansion using UMLS
OHSU-oshu_as_is_1000	AT	0.275	0.281	standard input with additional stop words

Our baseline textual run had a better than average performance, with a MAP of 0.28. The use of query expansion with UMLS synonyms as well as query parsing further improved the MAP. However, the most notable improvement was with the use of our modality classifier. By incorporating visual information, the MAP increases to 0.408, which is significantly better than any other official run submitted.

1.5 Conclusions and Future Work

Our image retrieval system built using open-source tools is a flexible platform for evaluating various tools and techniques in image processing as well as natural language processing for medical image retrieval. The use of visual information to automatically extract the imaging modality is a promising approach for the ImageCLEFmed campaign. The use of UMLS term expansion, query parsing and modality detection all add value over the basic Ferret (Lucene) search engine. We will continue to improve our image retrieval system by adding more image tags using automatic visual feature extraction. Our next goal is to annotate the images with their anatomical location and view attributes.

2 Automatic Image Annotation Task

The goal of this task was to correctly classify 1000 radiographic medical images using the hierarchical IRMA code. This code classifies the image along the modality, body

orientation, body region, and biological system axes. There were 116 unique classes. The task organizers provided a set of 9,000 *training* images and 1000 *development* images. The goal of the task was to classify the images to the most precise level possible, with a greater penalty applied for incorrect classification than for a less specific classification in the hierarchy.

2.1 Introduction

A supervised machine learning approach using global gist features and neural network architecture was employed for the task of automatic annotation of medical images with the IRMA code.

2.2 System Description

The automatic image annotation was based on a neural network classifier using Gist features [14]. The classifiers were created in MATLAB using the Netlab toolbox [15]. All images were convolved with a set of 32 multiscale-oriented Gabor filters. We created a 512-dimensional vector using statistics from these filters. Principal component analysis was then used to reduce the dimensionality of the vector to 100. A multilayer perceptron with one hidden layer containing 250-500 nodes was used to create and train a multi-class classifier. The training data set of 10,000 images was used to optimize performance of the development set of 1000 images. The final configuration of the classifier used 300 hidden nodes.

A confusion matrix was used to identify the most common mode of misclassification. We noted that classes 1123-110-500-000 (108) and 1123-127-500-000 (111) were frequently interchanged by our classifier. This error arises from the similarity of the Anterior-Posterior (AP) and the PA views of chest x-rays. To handle this special case, we created a second layer of classification built around a support vector machine (SVM) using scale-invariant feature transform (SIFT) features [16] as inputs. This new binary classify was used to determine the final class assignments for images in classes 108 and 111.

2.3 Runs Submitted

OHSU submitted two runs for the automatic annotation task. The first run used gist feature vectors to train the multi-layer perceptron. A neural network was used to create a multi-class classifier consisting of 116 classes. These were the original classes from 2006 and did not use the hierarchical nature of the IRMA code. These classes were then converted to the IRMA code, as required for the submission in 2007. The second run used a hierarchical classifier architecture, with the first layer as described above and the second classifier using SIFT features and an SVM.

2.4 Results and Analysis

The relationship between semantic and visual hierarchy remains an open area of research. Based on our experiments using this collection of images used for automatic annotation, the use of hierarchy of the semantic classes did not improve our automatic annotations as visual hierarchy did not correspond to semantic hierarchy.

The error count for both our runs were quite similar at 67.8 and 67.97 for 1000 images, compared to the best count of 26.84 and worst count of 505.61. There was only a very slight improvement in using the two-layer classifier. There were 227 errors using the 2006 classes, which corresponds to an classification accuracy of 77.3%. However, of these 227 errors, only 15 were wrong along all 4 axes. 76 were misclassified along two axes (primarily view and anatomy) while 12 were misclassified along 3 axes. 77 of our single misclassifications were along the view axis. A significant portion of these occurred where class 111 was misclassified as 108, an error due to confusion between posterior-anterior and anterior-posterior views of the chest.

2.5 Future Work

We would like to further investigate the mapping between the semantic and visual hierarchy of images in the IRMA collection. We primarily used a flat classifier in this work, with a constant cost for all classes and misclassifications. However, it might be possible to improve the performance using the IRMA hierarchy by the use of a cost function that depends on the hierarchy of the IRMA classes.

Acknowledgments

We acknowledge the support of NLM Training Grant 1T15 LM009461 and NSF Grant ITR-0325160. We would also like to thank Steven Bedrick for his help in creating the web-based image retrieval system.

References

1. Hersh, W., Muller, H., et al.: Advancing biomedical image retrieval: development and analysis of a test collection. *J. Am. Med. Inform. Assoc.* 13(5), 488–496 (2006)
2. Smeulders, A.W.M., Worring, M., et al.: Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380 (2000)
3. Tagare, H.D., Jaffe, C., et al.: Medical Image Databases: A Content-Based Retrieval Approach. *J. Am. Med. Inform. Assoc.* 4(3), 184–198 (1997)
4. Aisen, A.M., Broderick, L.S., et al.: Automated storage and retrieval of thin-section CT images to assist diagnosis: System description and preliminary assessment. *Radiology* 228, 265–270 (2003)
5. Schmid-Saugeon, P., Guillod, J., et al.: Towards a computer-aided diagnosis system for pigmented skin lesions. *Computerized Medical Imaging and Graphics* 27, 65–78 (2003)
6. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medicine – clinical benefits and future directions. *Int. J. Med. Inform.* 73, 1–23 (2004)
7. Hersh, W., Kalpathy-Cramer, J., et al.: Medical image retrieval and automated annotation: OHSU at ImageCLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006. LNCS*, vol. 4730, pp. 660–669. Springer, Heidelberg (2007)

8. Kalpathy-Cramer, J., Hersh, W.: Automatic Image Modality Based Classification and Annotation to Improve Medical Image Retrieval. In: MedInfo 2007, Brisbane, Australia, pp. 1334–1338 (2007)
9. Braschler, M., Peters, C.: Cross-language evaluation forum: objectives, results, achievements. *Inform Retrieval* (7), 7–31 (2004)
10. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the Image-CLEFmed 2006 medical retrieval annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)
11. Müller, H., Clough, P., et al.: Evaluation Axes for Medical Image Retrieval Systems - The ImageCLEF Experience. In: ACM Int. Conf. on Multimedia, Singapore (November 2005)
12. Florea, F., Müller, H., Rogozan, A., Geissbühler, A., Darmoni, S.: Medical image categorization with MedIC and MedGIFT. In: Medical Informatics Europe (MIE 2006) (2006)
13. Smith, L., Rindfleisch, T., Wilbur, W.: MedPost: a part-of-speech tagger for biomedical text *Bioinformatics*. 20(14) (2004)
14. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Computer Vision* 42(3), 145–175 (2001)
15. Nabney, I.T.: *Netlab: Algorithms for Pattern Recognition*. Springer, London (2004)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision* 60(2), 91–110 (2004)

Using Bayesian Network for Conceptual Indexing: Application to Medical Document Indexing with UMLS Metathesaurus

Thi Hoang Diem Le, 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
{stuthdl,viscjp,jooHwee}@i2r.a-star.edu.sg

Abstract. We describe a conceptual indexing method using UMLS meta-thesaurus. Concepts are automatically mapped from text using MetaMap software tool for English, and a simplified mapping tool for other languages. The concepts and their semantic links given by UMLS are used to build a Bayesian network. Retrieval process is then an inference process of probabilities or weights. Different types of relations are experimented in this model to evaluate their efficiency in retrieval.

1 Introduction

Conceptual indexing consists of the use of concepts to describe document index, instead of keywords. Concepts can be described as a unique and abstract human understandable notions independent from any direct material support, independent from any language or information representation, and are used to organize perception and knowledge [1]. As a consequence, an index using concepts is implicitly multi-lingual. The difficult part of this approach is the need of a prerequisite of an existing knowledge resource likely to be used during the automatic indexing process. Of course, building such a conceptual resource, that exhaustively describes a set of concepts for a given domain, is also a costly challenge. The second difficulty lies in the automatic mapping algorithm to associate correct concepts to sentences.

Since the beginning of Medical CLEF in 2005, we have explored with success the path of conceptual indexing, thanks to the availability of UMLS¹ and MetaMap software for identifying concepts from English text. With the use of conceptual structure of queries and documents, we have already shown that a conceptual indexing with an adapted knowledge base outperforms [1] classical text indexing. In this work, we focus on the *mismatch problem* that does appear when conceptual indexing takes place.

One may wonder why in conceptual domain, we may still have mismatch problems. In fact, using concepts instead of words or noun phrases for indexing solves

¹ <http://www.nlm.nih.gov/research/umls/>

several well known mismatch problems: homonymy, polysemy [2] and synonymy. Most of the synonymy problem [3,4] lies in the term variation phenomenon. If all synonym term variations can be replaced by a unique concept, then at the stage of conceptual indexing, the problem is solved. For example in UMLS², the term variations "skin disease", "skin disorder", "cutaneous disease", "cutaneous disorder", "dermatological disease", "dermatological disorder", etc, appear under the unique concept "C0037274". Unfortunately, in practice, there are still other problems to solve. As a term expressed in words may have intersection at character level, like "skin inflammation" and "skin disease", this syntactic link that reveals a semantic link, is lost when replacing these two terms by their counterpart concepts: "skin inflammation" with the concept "C0011603", and "skin disease" with the different concept "C0037274". In this experiment, we test the use of different semantic links between concepts in the matching process in order to solve this problem. We consider that this conceptual network forms a Bayesian network.

2 Bayesian Network Conceptual Indexing

The Bayesian network based model, is adapted from [5]. Our goal is to capture concepts, semantic relationships between document concepts, and query concepts in order to solve the concept mismatch problem. The retrieval process is tailored to exploit some probabilistic inferences from a document node to a query node. Initial probabilities of nodes are assigned in a similar way as weights in a usual retrieval system. Probability propagation is performed from a document to the query along the Bayesian network.

Conceptualization is the process of specifying what is meant by a term. In an IR context, it consists of the replacement of terms or noun phrases by a concept from an ontology. In our experiments, the conceptualization is based on the UMLS (Unified Medical Language System) meta-thesaurus. As this structure is not fully an ontology, because there is no formal description of concepts, it is quite limited for computing relation between concepts: UMLS just stores a large list of possible links between any two concepts. Moreover, there is no real guarantee of the actual quality of concepts stored in this structure. Finally, as UMLS is a fusion of different existing resources, relations between concepts are neither complete nor consistent. Despite all these problems, we still have interesting results using this structure compared to classical word based indexing method.

The method of concept extraction (conceptualization) is the same as our previous work in CLEF2006 [6], i.e. we use Metamap [7] for mapping concepts in English reports, and our tool for French and German reports after tagging by TreeTager. These concepts extracted are then used to build up the network for our Bayesian network based retrieval framework.

Bayesian network³ is a Direct Acyclic Graph (DAG) [8]. Our Bayesian network model includes document nodes, query nodes, concept nodes and direct

² Version 2007AC.

³ Called also belief network.

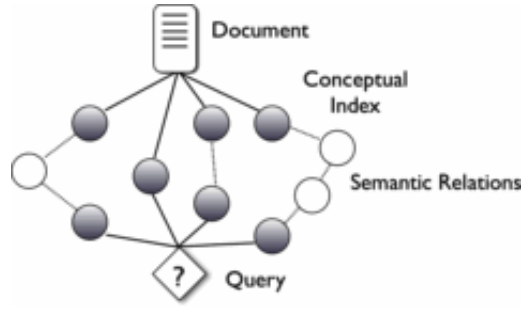


Fig. 1. Bayesian network on conceptual index

links between nodes (see Fig. 1). There are two types of links: links which connect concept nodes to documents or queries in which these concepts appear and links between document concept nodes and query concept nodes if there are semantic relations between them found in UMLS. The first type of link is in fact an indexing link, and the second one is semantic relation. These semantic relations are derived from UMLS relationship database files: MRREL.RRF for direct relation and MRHIER.RRF for hierarchical context of each instance of concept in different knowledge sources of UMLS. We decide to follow the direction of links from documents to queries so that we can calculate the probabilistic inference from documents to queries in the network. Nodes which are pointed by a link are considered as child nodes.

The retrieval process is a probabilistic inference from document node to query node, considered as a 3-steps process. These steps are explained below.

2.1 Prior Probability Initiation

First a document (for example D) is observed. Knowing this document, we have to compute the probability of a single concept to appear in this document: $P(c|D)$. We chose to estimate this prior probability for document concept nodes by a normalized term frequency divided by the inverse document frequency ($tf * idf$):

$$P(c|D) = \frac{w_c}{\sum_{c' \in D} w_{c'}} \text{ with } w_c = tf * idf \quad (1)$$

2.2 Inference from Document Concept Nodes to Query Concept Nodes

Prior probabilities of document concept nodes are then inferred to query concept nodes via semantic links between them by formula (2):

$$P(c|pa(c)) = \max_{pa_i(c) \in pa(c)} P(c|pa_i(c)) * P(pa_i(c)) \quad (2)$$

In this formula, $pa(c)$ is the set of parent concepts of c , $P(c|pa_i(c))$ is the relatedness between c and its parents according to the type of semantic link. It

can also be dependent on the quality of knowledge source which supplies the relationship information. The bigger the value of relatedness is, the more relevant the two concepts are. $P(c|pa_i(c))$ can be empirically predefined in the range $[0, 1]$ according to the semantic relation. It can be estimated using the semantic similarity formulas of Leacock-Chorodo [9] in case of indirect hierarchical relations:

$$sim(c_1, c_2) = \log \frac{2 * L}{l(c_1, c_2)} \quad (3)$$

In this similarity, L is the maximum length of the taxonomy and $l(c_1, c_2)$ is the minimum length of path from concept c_1 to concept c_2 .

2.3 Estimation of Relevance Status Value

Finally, the relevance status value (RSV) of a query Q corresponding to a document D is the conditional probability of query node Q with the condition that document node D is observed. In this case we use the inference mechanism as [8] in which the probability of a node given their parents is calculated by the link matrix. Among different types of link matrix formulas, we chose weighted sum link matrix formulas in order to take into account at the same time the inferred probability and weight of query concept nodes:

$$RSV(Q, D) = P(Q|pa(Q), D) = \frac{\sum_{c_i \in Q} w_{c_i} P(c_i)}{\sum_{c_i \in Q} w_{c_i}} \quad (4)$$

$P(c_i)$ is the probability of query concept node (in the set $pa(Q)$) inferred from document concept nodes of D ; w_i is the weight of query concept node correspond to query Q . This weight is also estimated by *tf.idf* normalized similarly to document concept nodes.

In order to enhance the retrieval results, the relevant status value in the ranking list is then re-weighted by semantic groups of concepts: RSV is multiplied with the number of concepts in the matched concept set that corresponds to the three important semantic groups: Anatomy, Pathology, and Modality. The purpose is to emphasize the importance of these three semantic groups considering the structure of queries. This method is proven efficient by our works in CLEF2006.

3 Results of Medical Runs

For the CLEF image medical 2007 collection [10], we analyse the new image sets(MyPacs and Endoscopic) the same way as the collection last year. Our experiments are based on predefined set of semantic relations in UMLS. We use isa-indirect relations that derived from MRHIER.RRF and other types of relations: parent-child (PAR-CHD), broader-narrower (BR-RN), similar or alike (RL), related or possibly synonym (RQ) extracted from MRREL.RRF. The value of relatedness of semantic relations in all of these submitted runs is predefined

in the range (0,1). In addition to these runs, we experiment another run in which instead of predefined values of relatedness, we used semantic similarity measurement of Leacock-Chorodo for isa-indirect relation (run 8). This method yeilds better result.

The MAP results (see Table 1) have shown that taking into account semantic relation in a Bayesian network (runs 2-8) can enhance the retrieval, compared to conceptual vector space model (run 1). Conceptual vector space model, is just a Vector Space Model (VSM) where vector dimensions are concepts instead of words, with $tf * idf$ weighting and cosine for matching function.

Moreover, method of mesuring semantic relatednes (run 8) can be used efficiently in our Bayesian model to solve the mistmatch gap between document and query.

Table 1. MAP results of CLEFMed2007

Id	Run type	Rel	source	ISA	PAR-CHD	BR-RN	RL	RQ	Map	R-prec
1	Base line								0.2684	0.2818
2	3 relations	REL		0.2	0.01		0.01		0.2750	0.2830
3	All relations	REL		0.2	0.01		0.01	0.001	0.2774	0.2869
4	Isa only	REL		0.1					0.2783	0.2824
5	Isa only	REL		0.2					0.2784	0.2800
6	Isa only	REL		0.3					0.2752	0.2778
7	Isa only	REL		0.4					0.2711	0.2746
8	Leacock-Chorodo	HIER							0.2787	0.2803

4 Conclusion

We have proposed a Bayesian model who takes into account the semantic relationship between documents concepts and query concepts in an unified framework. Experimentation shows that this model can enhance the VSM by adding the semantic relatedness between concepts. The ISA relation alone seems more effective than combined with other semantic relations. Hierarchical relation using Leacock-Chorodo weighting is appears to be most effective. Improvements on relationship weighting issue as well as performance of model will be studied further.

References

1. Chevallet, J.P., Lim, J.H., Le, T.H.D.: Domain knowledge conceptual inter-media indexing, application to multilingual multimedia medical reports. In: ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal (2007)
2. Krovetz, R.: Homonymy and polysemy in information retrieval. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, pp. 72–79. Association for Computational Linguistics, Morristown (1997)

3. Tzoukermann, E., Klavans, J.L., Jacquemin, C.: Effective use of natural language processing techniques for automatic conflation of multi-word terms: the role of derivational morphology, part of speech tagging, and shallow parsing. *SIGIR Forum* 31(SI), 148–155 (1997)
4. Jacquemin, C.: Syntagmatic and paradigmatic representations of term variation. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 341–348. Association for Computational Linguistics, Morristown (1999)
5. Turtle, H., Croft, W.B.: Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.* 9(3), 187–222 (1991)
6. Lacoste, C., Chevallet, J.P., Lim, J.H., Wei, X., Raccoceanu, D., Hoang, D.L.T., Teodorescu, R., Vuillenemot, N.: Ipal knowledge-based medical image retrieval in imageclefmed 2006. In: *Working Notes for the CLEF 2006 Workshop*, 20-22 September, Medical Image Track, Alicante, Spain (2006)
7. Aronson, A.R.: Metamap: Mapping text to the umls metathesaurus (2006), <http://mmtx.nlm.nih.gov/docs.shtml>
8. Pearl, J.: *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco (1988)
9. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) *WordNet: An Electronic Lexical Database*, pp. 265–284. MIT Press, Cambridge (1998)
10. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary (2007)

Baseline Results for the ImageCLEF 2007 Medical Automatic Annotation Task Using Global Image Features

Mark O. Güld and Thomas M. Deserno

Department of Medical Informatics, RWTH Aachen University, Aachen, Germany
`mgueldd@mi.rwth-aachen.de`, `deserno@ieee.org`

Abstract. This paper provides baseline results for the medical automatic annotation task of CLEF 2007 by applying the image retrieval in medical applications (IRMA)-based algorithms previously used in 2005 and 2006, with identical parameterization. Three classifiers based on global image features are combined within a nearest neighbor (NN) approach: texture histograms and two distance measures, which are applied on down-scaled versions of the original images and model common variabilities in the image data. According to the evaluation scheme introduced in 2007, which uses the hierarchical structure of the coding scheme for the categorization, the baseline classifier yields scores of 51.29 and 52.54 when reporting full codes for 1-NN and 5-NN, respectively. This corresponds to error rates of 20.0% and 18.0% (rank 18 among 68 runs), respectively. Improvements via addressing the code hierarchy were not obtained. However, comparing the baseline results yields that the 2007 task was slightly easier than the previous ones.

1 Introduction

The ImageCLEF medical automatic annotation task (MAAT) 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 MAAT [2] consisted of 10,000 reference images grouped into 116 categories and 1,000 images to be automatically categorized. The categorization is based on a medical code introduced in [3], which has since been refined and extended for new imaging techniques and body regions. In 2007, the hierarchical structure of the code is used to describe the image contents, with the evaluation scheme allowing a finer granularity of the classification accuracy [4]. Sets of 10,000 training images, 1,000 images for parameter optimization, and 1,000 unknown images are used in the experiments, again from 116 categories, i.e. unique codes. As both the task and the participants (i.e. classification algorithms) evolved over the last three years, it is difficult to compare the results. In this paper, we apply one algorithm with a fixed set of parameters to all the tasks, and therefore provide baseline results that allow a rough comparison between any pair of other runs over the last three years.

2 Methods

The image retrieval in medical applications (IRMA) framework is used to produce the baseline results [5]. In particular, the image content is represented by global features [6,7], where each image is assigned to one feature vector. Texture properties as proposed by TAMURA et al. are extracted for each image (scaled to 256×256 pixels), obtaining a 384-dimensional histogram. The distance between a pair of images is computed via the Jensen-Shannon divergence (JSD) of their respective texture histograms. Down-scaled representations of the images allow to explicitly model frequent, class-invariant variabilities among the images, such as translation, radiation dose, or local deformations. The down-scaled images are compressed to roughly 1KB of size, 32×32 pixels when applying the cross-correlation function (CCF) as a similarity measure, or gradient images of $X \times 32$ pixels when applying the image distortion model (IDM), regarding and acknowledging the original aspect ratio, respectively. The combination of these features is done within a NN scheme: a total distance between a sample image q and a reference image r is obtained by the weighted sum of the normalized distances from the single classifiers. The weighting coefficients were empirically adjusted based on prior (non-CLEF MAAT) experiments.

In 2007, the evaluation is done using the scheme described in [4]. For each image from the test set, an error value $e \in [0..1]$ is obtained, based on the position of classification errors in the hierarchy. By summation over all 1,000 test images, the overall value is obtained. Constantly answering *don't know* yields a value of 500.0, the worst possible value is 1000.0. To address the modified evaluation scheme, the NN decision rule is modified in an additional experiment: From the k neighbors, a *common* code is generated by setting differing parts (and their subparts) to *don't know*, e.g. two neighbors with codes 1121-120-434-700 and 1121-12f-466-700 result in a *common* code of 1121-12*-4**-700.

3 Results

All results are obtained non-interactively, i.e. without relevance feedback from a human. Tab.1 contains the baseline error rates. In 2007, the evaluation was not based on the error rate – the table also contains the rank based on the modified evaluation scheme for the corresponding submission of full codes. Runs which were not submitted are displayed marked with asterisks, along with their hypothetical rank. Using the evaluation scheme proposed for 2007, the default weighting for k -NN yields 51.29 and 52.54 for $k = 1$ and $k = 5$, respectively. The *common code* rule yields 80.47 when applied to the 5-NN results.

4 Discussion

The medical automatic annotation task in 2007 is a bit easier than 2006, as the baseline error rate drops from 21.7% to 20.0% and from 22.0% to 18.0% for the 1-NN and the 5-NN, respectively. As the baseline results are reported for

Table 1. Baseline error rates (ER) and ranks among submissions

Year	References	Classes	$k = 1$		$k = 5$	
			ER	Rank	ER	Rank
2005	9,000	57	13.3%	2/42	14.8%	*7/42
2006	10,000	116	21.7%	13/28	22.0%	*13/28
2007	11,000	116	20.0%	*17/68	18.0%	18/68

the last three years, they allow the a rough comparison between submissions from the years 2005, 2006, and 2007 for algorithms which only participated in one year. Note that the evaluation scheme significantly differs from the error rate: Although the 1-NN yields a better result than the 5-NN, its error rate is actually worse. This depends on the severity of misclassifications, which are captured by the new evaluation scheme.

The *common code* rule, which generates a code fragment that all k nearest neighbors agree on, does not improve the results, but performs significantly worse. In addition, other experiments were performed to either remove neighbors from the NN list by applying a distance threshold, or by modifying the complete decision into *don't know*, again based on a distance threshold. These experiments did not provide any improvement, either. This seems to be consistent with efforts by other groups, which were largely unable to improve their results if they address the code hierarchy.

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., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)
3. 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. 109–117 (2003)
4. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)

5. Güld, M.O., Thies, C., Fischer, B., Lehmann, T.M.: A generic concept for the implementation of medical image retrieval systems. *International Journal of Medical Informatics* 76(2-3), 252–259 (2007)
6. Lehmann, T.M., Güld, M.O., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohlen, M., Schubert, H., Wein, B.B.: Content-based image retrieval in medical applications. *Methods of Information in Medicine* 43(4), 354–361 (2004)
7. 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)

Evaluation of Automatically Assigned MeSH Terms for Retrieval of Medical Images

Miguel E. Ruiz¹ and Aurélie Névéol²

¹ University of North Texas, School of Library and Information Sciences
P.O. Box 311068, Denton, Texas 76203-1068 USA
meruiz@unt.edu

² National Library of Medicine
Bldg. 38A, B1N-28A, 8600 Rockville Pike, Bethesda, MD 20894 USA
neveola@mail.nih.gov

Abstract. This paper presents the results of the State University of New York at Buffalo (UB) team in collaboration with the National Library of Medicine (NLM) in the 2007 ImageCLEFmed task. We use a system that combines visual features (using a CBIR System) and text retrieval. We used the Medical Text Indexer (MTI) developed by NLM to automatically assign MeSH terms and UMLS concepts to the English free text annotations of the images. We also used an equivalent system called MAIF that automatically assigns MeSH and UMLS concepts to French free text. Our results indicate that the use of automatically assigned UMLS concepts improves retrieval performance significantly. We also identified specific aspects of the system that could be improved in the future, such as the method used to perform the automatic translation of medical terms and the addition of image classification to process queries targeted to a specific image modality.

1 Introduction

This paper presents the results of our participation in imageCLEFmed 2007. In previous years we have used a method that maps the queries to Unified Medical Language System (UMLS) concepts and then uses these concepts to find translations of the English queries into French and German [1, 2]. This method has been successful in handling English queries to find the corresponding French and German translations.

For this year's challenge, we focused on assessing 1) the use of an automatic indexing system providing Medical subject Headings (MeSH terms) and UMLS concepts; and 2) the use of UMLS-based translation with French as the query language. The impact of both features on retrieval performance was analyzed.

2 System Description

The system that was used this year combines two publicly available systems:

- SMART: This is an information retrieval system developed by Gerald Salton and his collaborators at Cornell University [3]. SMART implements a generalized vector space model representation of documents and queries.

This is an important feature since we wanted to include three different representations of the image annotations: Free text, MeSH terms, and UMLS concepts.

- Flexible Image Retrieval Engine (FIRE): This is an open source content based image retrieval system developed at RWTH Aachen University, Germany [4].

For processing the annotations we also used two automatic text categorization tools that map free text to MeSH terms. We used the Medical Text Indexer (MTI) which is a tool developed at the U.S. National Library of Medicine (NLM) to assign MeSH terms to the English annotations. For processing French text we used Medical Automatic Indexer for French (MAIF) which is a tool similar to MTI that uses NLP as well as statistical methods to assign MeSH terms to free text. We did not have a tool to perform a similar mapping of the German text.

We also decided to add the concept unique identifier (CUI) from the UMLS so that we could match queries and documents using these language independent concepts. Since MeSH is one of the vocabularies of UMLS, the assignment of the UMLS concepts was performed by getting the corresponding identifiers of the MeSH terms in UMLS.

3 Collection Preparation

As described in the ImageCLEFmed 2007 overview paper [5] the image collection used in this task consists of six sub-collections. Each collection has its own meta-data in XML format for the image annotations. In order to process all collections uniformly we created a common XML schema and converted all the annotation to this new schema. Figure 1 shows the common metadata schema that was used.

English queries and documents were processed by parsing them using MTI to identify MeSH concepts present in the free text and then add the corresponding MeSH terms as well as the UMLS concepts. MTI uses NLP techniques (implemented in Metamap) as well as a statistical K-Nearest-Neighbor (KNN) method that takes advantage of the entire MEDLINE collection [6]. MTI is currently being used at NLM as a semi-automatic and fully automatic indexing tool. For this task, we used the top 25 recommendations provided by the system ran with default filtering.

French queries and documents were processed using a modified version of the MAIF described in [7]. MAIF is able to retrieve MeSH terms from biomedical text in French. It specifically retrieves main headings and main heading/subheading pairs. However, for the purpose of the image-CLEF task, we only used MAIF to retrieve MeSH main headings that were then mapped to UMLS concepts. We used a collection of 15,000 French citations available from CISMeF (Catalogue and Index of Online Health Information in French available at www.cismef.org) for retrieving the French MeSH terms used in MAIF. The modified version of MAIF is similar to MTI in that it combines a NLP method and a statistical, knowledge-based method [7]. However, the two systems differ in the specific implementation of both methods. The combination of these two approaches takes

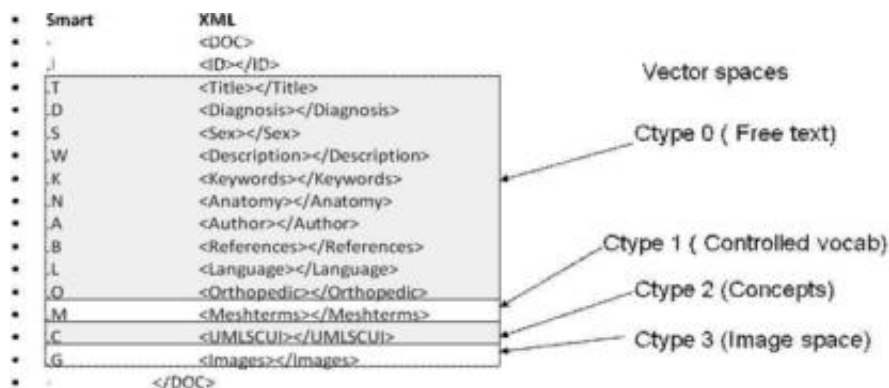


Fig. 1. Common XML schema and Ctypes for indexing

into account the relative score assigned to the terms by each method. The "relative score" of a term is obtained by dividing the score of the term by the sum of all the scores assigned by the corresponding method. Combining the methods in this way gives an advantage to terms retrieved by the NLP method. Because the NLP approach tends to retrieve a smaller number of terms per document, the relative importance of each term tends to be higher than the relative importance of terms retrieved by the statistical method. The final term selection is performed using the breakage function described in [8]. The score assigned to a MeSH candidate represents its likelihood to be a good indexing term: the higher the score, the more likely it is that the corresponding MeSH term is a good indexing candidate. Given a list of indexing candidates and the score that has been assigned to them, the breakage function is meant to detect a breach of continuity in the scores, therefore highlighting the point in the candidate list where terms become significantly less likely to be correct indexing terms. The final set of MeSH main headings assigned to a document consists of all the terms ranked above this threshold.

Once the collections were converted in to the common XML schema we use SMART to parse the XML documents and create three indexes (also called Ctypes in SMART). Ctype 0 was used for indexing free text from the original annotations, Ctype 1 was used to index the MeSH terms automatically assigned using the medical text indexing tools (MTI for English text and MAIF for French text), and Ctype 2 was used to index the UMLS concepts that were identified by MTI or MAIF.

4 Retrieval Model

We used a generalized vector space model that combines the vector representation of each of the four indexes presented in Figure 1. The final retrieval model can be represented using the following formula:

$$score(image) = \alpha * Score_{CBIR} + \beta * sim_{Text}(d_i, q) \quad (1)$$

where α and β are coefficients that weight the contribution of each system and sim_{Text} is defined as:

$$sim_{Text}(d_i, q) = \lambda * sim_{words}(d_i, q) + \mu * sim_{MeSHterms}(d_i, q) + \rho * sim_{UMLSConcepts}(d_i, q) \quad (2)$$

where λ , μ and ρ are coefficients that control the contribution of each of the ctypes. The values of these coefficients were computed empirically using the optimal results on the 2006 topics. The similarity values are computed using cosine normalization (*atc*) for the documents and augmented term frequency for the queries (*atn*). We also performed automatic retrieval feedback by retrieving 1,000 documents using the original query and assuming that the top n documents are relevant. This allowed us to select the top m terms ranked according to Rocchio's relevance feedback formula [9].

5 Experimental Results and Analysis

We submitted 7 official runs which are shown in Table 1. A total of 5 runs use queries in English and 2 runs use queries in French. Translations of the queries into the other two languages were automatically generated by expanding the query with the all UMLS terms associated to the concepts assigned by MTI or MAIF. From these runs we can see that the highest score was obtained by runs that use the English queries and combine the text and image results obtaining a Mean Average Precision (MAP) value of 0.2938 and 0.293 (UB-NLM-UBTL3, and UB-NLM-UBTL1). Overall these two runs perform well above the median run in imageCLEFmed 2007 (Median MAP= 0.1828) and rank 5th and 6th among all automatic mixed runs. Unfortunately our multilingual runs perform significantly below (MAP 0.254). This indicates that our automatic translation approach does decrease performance when compared to using the English queries only. We suspect that this could be due to the fact that the translations might be adding terms that change the focus of the query.

Tables 2a-d show a series of unofficial runs that allow comparison of the methods that were used in our system. Table 2a shows the performance obtained by using free text (English only), automatically assigned UMLS concepts and the CBIR retrieval using FIRE. Our base lines for free text and UMLS concepts are quite strong since they both perform above the median system. The CBIR baseline is quite weak compared with the text and concept baselines. However, when compared to other visual only runs it is around average for CBIR runs. A query by query analysis of the results for the CBIR run shows that the MAP for 21 of the 30 queries is below 0.0001 which is a major factor for the poor performance shown. It appears that the fact that the queries require specific image modality seems to be a major factor since our CBIR system does not include an image classification module that could identify the image modality to filter out those images that do not have the requested modality in the query.

Table 2b shows the results obtained using only English queries. Because the collection has predominantly English annotations we can see that these runs

Table 1. Performance of Official Runs

Run name	Description	type	MAP	Exact-P	P10	P20
UB-NLM-UBTL3	English queries	Mixed run	0.2938	0.2893	0.4167	0.3867
UB-NLM-UBTL1	English queries	Mixed run	0.293	0.2992	0.4000	0.3933
UB-NLM-UBmixedMulti2	English cross-lang	Mixed run	0.2537	0.2579	0.3167	0.3017
UB-NLM-UBTextBL1	English queries	Text only	0.2833	0.2833	0.4100	0.3817
UB-NLM-UBTextBL2	English cross-lang	Text only	0.2436	0.2461	0.3033	0.3017
UB-NLM-UBTextFR	French cross-lang	Text only	0.1414	0.1477	0.1933	0.1650
UB-NLM-UBmixedFR	French cross-lang	Mixed run	0.1364	0.1732	0.2000	0.1933

correspond to our highest scoring official runs (UBTL1 and UBTL3). All these runs use the free text as well as the UMLS concepts automatically assigned to both queries and documents. These results confirm that the use of automatically identified concepts improves performance considerably when compared to using free text only. We can also see that the merging formula that combines visual and text features does work properly despite the fact that the CBIR run contributes little to the overall MAP. Our two top scoring runs use text as well as image features. The best automatic run (MAP=0.3018) was not submitted but is only marginally better than our highest official run. Table 2c and 2d show performance of our cross-lingual runs. These runs use the UMLS automatic translations based on the UMLS concept mapping obtained from the English text. We can see that this actually harms performance significantly compared with using English only queries. We believe that is due to the aggressive translation method that we tried to use since it seems to add terms that shift the focus of the query. We plan to explore this issue in more detail in our future research. Despite this result we can see that the results confirm that using UMLS concepts (which are language independent) improves performance with respect to using only free text translations. Also the use of the results from the CBIR system yield only small improvements in retrieval performance.

Table 2d shows the result of our cross-lingual runs that use French as the query language. Our official French runs used the same parameters as the English runs and this seems to have harmed the results for French since the runs presented in our unofficial runs show significantly better performance. These results are comparable to the best French cross-lingual results presented by other teams in the conference. However, the overall French cross-lingual results achieve only 56% of the English retrieval performance. This could be due to the fact that the French resources we used (citation database and medical lexicon) are much smaller than the UMLS resources available for English.

Table 3 presents runs that use all the manually generated terms in English, French and German that were provided in the ImageCLEFmed topics. These queries achieve the highest score using our system with a MAP of 0.3148 which is comparable to the best manual run reported this year [5]. As in our previously presented experiments, the results with the manual queries show improvements when automatically generated UMLS concepts and pseudo relevance feedback are used. Use of the CBIR results yields a small improvement.

Table 2. Unofficial Runs

Run name	MAP	Exact-P	P10	P20
(a) Baseline runs				
EN-free text only	0.2566	0.2724	0.4000	0.3433
UMLS concepts only	0.1841	0.2007	0.2655	0.2345
FIRE baseline (CBIR)	0.0096	0.0287	0.0300	0.0317
(b) English only runs				
EN-text-RF	0.2966	0.2782	0.4033	0.3800
EN-text baseline + image	0.2965	0.3189	0.4067	0.3817
EN-text rf + images	0.3028	0.2908	0.4033	0.3800
(c) Automatic English cross-lingual runs				
EN-Multi-Baseline	0.2111	0.2283	0.2533	0.2467
EN-Multi + concepts	0.2789	0.2975	0.3400	0.3100
EN-Multi + concepts + images	0.2800	0.2981	0.3433	0.3117
EN-Multi-rf	0.2789	0.2975	0.3400	0.3100
(d) Automatic French cross-lingual runs				
FR-Multi-Baseline	0.1442	0.1456	0.1700	0.1500
FR-Multi-Baseline + images	0.1453	0.1466	0.1700	0.1550
FR-Multi- RF	0.1618	0.1680	0.2133	0.1883
FR-Multi-RF + images	0.1707	0.1873	0.2167	0.1967

Table 3. Manual runs

Run name	MAP	Exact-P	P10	P20
Multi-manual text only	0.2655	0.3082	0.3933	0.3467
Multi-Manual text+contepts	0.3052	0.3127	0.4133	0.3933
Multi-Manual Text+concepts + images	0.3069	0.3148	0.4167	0.3933
Multi-manual rf	0.3092	0.2940	0.4233	0.3983
Multi-manual rf + images	0.3148	0.3005	0.4200	0.3967

Table 4. Comparison of results by type of query

Type	Free text	UMLS concepts	CBIR	Combination
Visual	0.19737	0.11478	0.01159	0.22064
Visual-Semantic	0.11375	0.1056	0.01508	0.20118
Semantic	0.32208	0.32275	0.00209	0.4596

We performed a query by query analysis to try to understand how the different methods proposed are affected by different types of queries. Table 4 shows the average MAP by groups of topics according to whether they are visual, semantic and mixed (visual-semantic). As expected the text based and UMLS concept based runs perform better in the semantic topics. The CBIR system performs slightly better in the visual and mixed topics while the poorest performance is

in the semantic topics. The combination shows consistent improvements in all three groups of topics.

6 Conclusions

From the results we can conclude that the use of automatically assigned UMLS concepts using MTI significantly improves performance for the retrieval of medical images with English annotations. We also confirm that our generalized vector space model works well for combining retrieval results from free text, UMLS concepts and CBIR systems. Despite the low performance of our CBIR system the merging method is robust enough to maintain or even improve results.

We also conclude that our methods work better for semantic queries while still achieving significantly high performance for visual or mixed visual semantic queries.

Our cross-lingual results using French as the query language are relatively low and indicate that we need to work on improving our translation method based on UMLS mapping. We plan to explore this further in our future research.

The low results from the CBIR system indicate that we need to address the image classification problem so that the CBIR results can give a more significant contribution to the overall fusion of results.

Acknowledgements

This work was supported in part by an appointment of A. Név  ol and M. E. Ruiz to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

We also want to thank Dr. Alan Aronson and the Indexing Initiative Project team at the NLM for their support and for making the MTI system available for this project.

References

- [1] Ruiz, M.: Combining image features, case descriptions and umls concepts to improve retrieval of medical images. In: Proceedings of the AMIA Annual Symposium, Washington, DC, pp. 674–678 (2006)
- [2] Ruiz, M.: Ub at imageclefmed 2006. In: Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 702–705. Springer, Heidelberg (2007)
- [3] Salton, G. (ed.): The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs (1983)
- [4] Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: A quantitative comparison. In: Rasmussen, C.E., B  lthoff, H.H., Sch  lkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 228–236. Springer, Heidelberg (2004)

- [5] Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Hersh, W.: Overview of the imageclef 2007 medical retrieval and annotation tasks. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
- [6] Aronson, A., Mork, J., Gay, C., Humphrey, S., Rogers, W.: The.nlm indexing initiative's medical text indexer. In: MEDINFO, 11(Pt 1), pp. 268–272 (2004)
- [7] Névéal, A., Mork, J., Aronson, A., Darmoni, S.: Evaluation of french and english mesh indexing systems with a parallel corpus. In: Proceedings of the AMIA Annual Symposium, pp. 565–569 (2005)
- [8] Névéal, A., Rogozan, A., Darmoni, S.: Automatic indexing of online health resources for a french quality controlled gateway. *Information Processing and Management* 42, 695–709 (2006)
- [9] 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)

University and Hospitals of Geneva Participating at ImageCLEF 2007

Xin Zhou¹, Julien Gobeill¹, Patrick Ruch¹, and Henning Müller^{1,2}

¹ Medical Informatics Service, University and Hospitals of Geneva, Switzerland

² Business Information Systems, University of Applied Sciences, Sierre, Switzerland
`xin.zhou@sim.hcuge.ch`

Abstract. This article describes the participation of the University and Hospitals of Geneva at three tasks of the 2007 ImageCLEF image retrieval benchmark. The visual retrieval techniques relied mainly on the GNU Image Finding Tool (GIFT) whereas multilingual text retrieval was performed by mapping the full text documents and the queries in a variety of languages onto MeSH (Medical Subject Headings) terms, using the EasyIR text retrieval engine for retrieval.

For the visual tasks it becomes clear that the baseline GIFT runs do not have the same performance as more sophisticated techniques such as visual patch histograms do have. GIFT can be seen as a baseline for the visual retrieval as it has been used for the past four years in ImageCLEF in the same configuration. Whereas in 2004 the performance of GIFT was among the best systems it now is towards the lower end, showing the clear improvement in retrieval quality. Due to time constraints no further optimizations could be performed and no relevance feedback was used, one of the strong points of GIFT. The text retrieval runs have a good performance showing the effectiveness of the approach to map terms onto an ontology. Mixed runs are in performance slightly lower than the best text results alone, meaning that more care needs to be taken in combining runs. English is by far the language with the best results; even a mixed run of the three languages was lower in performance.

1 Introduction

ImageCLEF¹ [1] has started within CLEF² (Cross Language Evaluation Forum, [2]) in 2003 with the goal to benchmark image retrieval in multilingual document collections. A medical image retrieval task³ was added in 2004 to explore domain-specific multilingual information retrieval and also multi modal retrieval by combining visual and textual features for retrieval. Since 2005, a medical retrieval and a medical image annotation task are both presented as part of ImageCLEF [3].

More about the ImageCLEF tasks, topics, and results in 2007 can also be read in [4,5,6].

¹ <http://www.imageclef.org/>

² <http://www.clef-campaign.org/>

³ <http://ir.ohsu.edu/image/>

2 Retrieval Strategies

This section describes the basic technologies that are used for the retrieval. More details on small optimizations per task are given in the results section.

2.1 Text Retrieval Approach

The text retrieval approach used in 2007 is similar to the techniques already applied in 2006 [7]. The full text of the documents in the collection and of the queries were mapped to a fixed number of MeSH terms, and retrieval was then performed in the MeSH-term space. Based on the results of 2006, when 3, 5, and 8 terms were extracted we increased the number of terms further. It was shown in 2006 that a larger number of terms lead to better results, although several of the terms might be incorrect, these incorrect terms create less damage than the few additionally correct terms add in quality. Thus 15 terms were generated for each document in 2007 and 3 terms from every query, separated by language. Term generation is based on a MeSH categorizer [8,9] developed in Geneva. As MeSH exists in English, German, and French, multilingual treatment of the entire collection is thus possible. For ease of computation an English stemmer was used on the collection and all XML tags in the documents were removed, basically removing all structure of the documents. The entire text collection was indexed with the EasyIR toolkit [10] using a pivoted-normalization weighting schema. Schema tuning was discarded due to the lack of time.

Queries were executed in each of the three languages separately and an additional run combined the results of the three languages.

2.2 Visual Retrieval Techniques

The technology used for the visual retrieval of images is mainly taken from the *Viper*⁴ project [11]. An outcome of the *Viper* project is the *GIFT*⁵. This tool is open source and can be used by other participants of ImageCLEF. A ranked list of visually similar images for every query topic was made available for participants and serves as a baseline to measure the quality of submissions. Feature sets used by *GIFT* are:

- Local color features at different scales by partitioning the images successively into four equally sized regions (four times) and taking the mode color of each region as a descriptor;
- global color features in the form of a color histogram, compared by a simple histogram intersection;
- local texture features by partitioning the image and applying Gabor filters in various scales and directions, quantized into 10 strengths;
- global texture features represented as a simple histogram of responses of the local Gabor filters at the smallest size in various directions and scales.

⁴ <http://viper.unige.ch/>

⁵ <http://www.gnu.org/software/gift/>

A particularity of *GIFT* is that it uses many techniques well-known from text retrieval. Visual features are quantized and the feature space is similar to the distribution of words in texts. A simple *tf/idf* weighting is used and the query weights are normalized by the results of the query itself. The histogram features are compared based on a histogram intersection [12].

3 Results

This section details the results obtained for the various tasks. It always compares our results to the best results in the competition to underline the fact that our results are a baseline for comparison of techniques.

3.1 Photographic Image Retrieval

The two runs submitted for the photographic retrieval task do not contain any optimizations and are a simple baseline using the GIFT system to compare the improvement of participants over the years. Only visual retrieval was attempted and no text was used. The two runs are fully automatic.

Table 1. Our two runs for the photographic retrieval task

run ID	MAP	P10	P30	Relevant retrieved
best visual run	0.1890	0.4700	0.2922	1708
GE_GIFT18_3	0.0222	0.0983	0.0622	719
GE_GIFT9_2	0.0212	0.0800	0.0594	785

Table 1 shows the results of the two submitted runs with GIFT compared to the best overall visual run submitted. MAP is much lower than the best run, almost by a factor of ten, whereas early precision is about a factor of five lower. The best run uses the standard GIFT system whereas the second run uses a smaller number of colors (9 hues instead of 18) and a smaller number of saturations as well. The results with these changes are slightly lower but the number of relevant images found is significantly higher, meaning that more fuzziness in the feature space is better for finding relevant images but less good concerning early precision.

3.2 Medical Image Retrieval

This section describes the three categories of runs that were submitted for the medical retrieval task (visual, textual, mixed). All runs were automatic and so the results are classified by the media used.

Visual Retrieval. The purely visual retrieval was performed with the standard GIFT system using 4 gray levels and with a modified gift using 8 gray levels. A third run was created by a linear combination of the two previous runs.

Table 2. Results for purely visual retrieval at the medical retrieval task

Run	num_ret	num_rel_ret	MAP	R-prec	bpref	P10	P30
best visual run	30000	1376	0.2427	0.264	0.283	0.48	0.3756
GE_4_8	30000	245	0.0035	0.0144	0.0241	0.04	0.0233
GE_GIFT8	30000	245	0.0035	0.0143	0.024	0.04	0.0233
GE_GIFT4	30000	244	0.0035	0.0144	0.024	0.04	0.0233

Table 2 shows the results of the best overall visual run and all of our runs. It is actually interesting to see that all but three visual runs have very low performance in 2007. These three runs used training data on almost the same collection of the years 2005 and 2006 to select and weight features, leading to an extreme increase in retrieval performance. Our runs are on the lower end of the spectrum concerning MAP but very close to other visual runs. Early precision becomes slightly better in the combination runs using a combination of two gray level quantizations.

Textual Retrieval. Textual retrieval was performed using each of the query languages separately and in addition in a combined run.

Table 3. Results for purely textual retrieval

Run	num_ret	num_rel_ret	MAP	R-prec	bpref	P10	P30
best textual run	28537	1904	0.3538	0.3643	0.3954	0.43	0.3844
GE_EN	27765	1839	0.2369	0.2537	0.2867	0.3333	0.2678
GE_MIX	30000	1806	0.2186	0.2296	0.2566	0.2967	0.2622
GE_DE	26200	1166	0.1433	0.1579	0.209	0.2	0.15
GE_FR	29965	1139	0.115	0.1276	0.1503	0.1267	0.1289

Results of our four runs can be seen in Table 3. The results show clearly that English obtains the best performance among the three languages. This can be explained as the majority of the documents are in English and the majority of relevance judges are also native English speakers creating both a potential bias towards relevant documents in English. For most of the best performing runs it is not clear whether they use a single language or a mix of languages, which is not really a realistic scenario for multilingual retrieval. Both, German and French retrieval have a lower performance than English and the run linearly combining the three languages is also lower in performance than English alone. In comparison to the best overall runs our system is close in number of relevant items found and still among the better systems in all other categories.

Mixed-Media Retrieval. There were two different sorts of mixed media runs in 2007 from the University and Hospitals of Geneva. One was a combination of our own visual and textual runs and the other was a combination of the GIFT results with results from the FIRE (Flexible Image Retrieval Engine) system and

a system from OHSU (Oregon Health and Science University). In these runs we discovered a problem we had with the evaluation of the treceval package that does not take into account the order of the items in the submitted runs. Some runs assumed the order to be the main criterion and had same weightings for many items. This can result in very different scores and for this reason we add in this table a recalculated map where the score is simply set to $1/\text{rank}$.

Table 4. Results for the combined media runs

Run	num_ret	num_rel_ret	MAP	new MAP	R-prec	bpref	P10	P30
best mixed run	21868	1778	0.3415	0.4084	0.3808	0.4099	0.4333	0.37
GE_VT1_4	30000	1806	0.2195	0.2199	0.2307	0.2567	0.3033	0.2622
GE_VT1_8	30000	1806	0.2195	0.2204	0.2307	0.2566	0.3033	0.2622
GE_VT5_4	30000	1562	0.2082	0.2090	0.2328	0.2423	0.2967	0.2611
GE_VT5_8	30000	1565	0.2082	0.2082	0.2327	0.2424	0.2967	0.2611
GE_VT10_4	30000	1192	0.1828	0.1829	0.2125	0.2141	0.31	0.2633
GE_VT10_8	30000	1196	0.1828	0.1839	0.2122	0.214	0.31	0.2633
3gift-3fire-4ohsu	29651	1748	0.0288	0.1564	0.0185	0.1247	0.0067	0.0111
4gift-4fire-2ohsu	29651	1766	0.0284	0.2194	0.0135	0.1176	0.0233	0.0156
1gift-1fire-8ohsu	29709	1317	0.0197	0.0698	0.0184	0.1111	0.0067	0.0133
3gift-7ohsu	29945	1311	0.0169	0.1081	0.0108	0.1309	0.0033	0.0044
5gift-5ohsu	29945	1317	0.0153	0.1867	0.0057	0.1151	0.0033	0.0022
7gift-3ohsu	29945	1319	0.0148	0.2652	0.0042	0.1033	0.0033	0.0022

The combinations of our visual with our own English retrieval run were all better in quality than the combinations with the FIRE and OHSU runs in the initial results but when re-scoring the images taking into account the rank information this changes completely! Combinations are all simple, linear combinations with a percentage of 10%, 50% and 90% of the visual runs. It shows that the smallest proportion of visual influence delivers the best results concerning MAP, although not as high as the purely textual run alone. Concerning early precision the runs with a higher visual proportion are on the other hand better than with a lower percentage. Differences between the two gray level quantizations (8 and 4) are extremely small.

3.3 Medical Image Classification

For medical image classification the basic GIFT system was used as a baseline for classification. It shows as already in [13] that the features are not too well suited for image classification as they do not include any invariance and are on a very low semantic level. Performance as shown in Table 5 is low compared to the best systems for our runs submitted for the competition.

The strategy was to perform the classification in an image retrieval way. No training phase was carried out. Visually similar images with known classes were used to classify images from the test set. In practice, the first 10 retrieved images

Table 5. Results of the runs submitted to the medical image annotation task

run ID	score
best system	26.847
GE_GIFT10_0.5ve	375.720
GE_GIFT10_0.15vs	390.291
GE_GIFT10_0.66vd	391.024

of every image of the test set were taken into account, and the scores of these images were used to choose the IRMA code on all hierarchy levels. When the sum of the scores for a certain code reaches a fixed threshold, an agreement can be assumed for this level. This allows the classification to be performed up to this level. Otherwise, this level and all further levels were not classified and left empty. This is similar to a classical kNN (k Nearest Neighbors) approach.

Thresholds and voting strategies varied slightly. Three voting strategies were used:

- Every retrieved image votes equally. A code at a certain level will be chosen only if more than half of the results are in agreement.
- Retrieved images vote with decreasing importance values (from 10 to 1) according to their rank. A code at a certain level will be chosen if more than 66% of the maximum was reached for a code.
- The retrieved images vote with their absolute similarity value. A code at a certain level will be chosen if the average of the similarity score for this code is higher than a fixed value.

Results in Table 5 show that the easiest method gives the best results. It can be concluded that a high similarity score is not a significant parameter to classify images.

New Runs. Based on our first experiences with the classification, several parameters were tried out to optimize performance without learning for the existing system. One clear idea was that taking only the first ten images was not enough, so up to the first 100 images were taken into account. The threshold was also regarded as too high favoring non-classification over taking chances. Another approach was to classify images not only on the entire hierarchy but also fixed on a full axis level or fixed for the entire code. In the competition the best systems did not take into account the hierarchy at all. Adding a simple aspect ration as feature further improved our results significantly (reduction in error score of around 100). All this brought down the overall classification to 234 instead of an initial 391, which is an enormous gain. Table 6 details the best results obtained with these changes. The best run actually performs classification on an axis-bases, thus takes into account part of the hierarchy.

Despite the enormous improvements in the error score it can clearly be seen that new feature sets and a learning strategy still have an strong potential for our approach.

Table 6. Results of some new runs to search for the optimization

run ID	score
GE_GIFT13_0.4vad_withAR	234.469972
GE_GIFT11_0.4vae_withAR	238.0446107
GE_GIFT100_vakNN_withAR	262.249183

4 Discussion

The results show clearly that visual retrieval with the GIFT is not state of the art anymore and that more specific techniques can receive much better retrieval results than a very simple and general retrieval system that did perform well in benchmarks three years ago. Still, the GIFT runs serve as a baseline as they can be reproduced easily as the software is open source and they have been used in ImageCLEF since 2004, which clearly shows the improvement of techniques participating in ImageCLEF since this time.

The text retrieval approach shows that the extraction of MeSH terms from documents and queries and then performing retrieval based on these terms is working well. Bias is towards the English terms with a majority of documents being in English and also the relevance judges being all native English speakers. In a truly multilingual setting with unbiased relevance judges, such an approach to map terms onto an ontology should even perform much better than the other approaches mixing languages.

Combining visual and textual retrieval remains difficult and in our case no result is as good as the English text results alone. Only early precision could be improved in visual retrieval. Much potential still seems to be in this combination of media.

For the classification of images our extremely easy approach was mainly hindered by the simple base features that were used and the absence of using the training data for optimization. Simple improvements such as the use of the aspect ration and slightly modified voting schemes improved the results already enormously.

Acknowledgements

This study was partially supported by the Swiss National Science Foundation (Grants 3200-065228 and 205321-109304/1) and the European Union (SemanticMining Network of Excellence, INFS-CT-2004-507505) via OFES Grant (No 03.0399).

References

1. Clough, P., Müller, H., Sanderson, M.: The CLEF cross-language image retrieval track (ImageCLEF) 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)

2. Savoy, J.: Report on CLEF-2001 experiments. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 27–43. Springer, Heidelberg (2002)
3. 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 (September 2006)
4. Deselaers, T., Hanbury, A., et al.: Overview of the ImageCLEF 2007 object retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
5. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
6. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
7. Gobeill, J., Müller, H., Ruch, P.: Translation by text categorization: Medical image retrieval in ImageCLEFmed 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
8. Ruch, P.: Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics* 22(6), 658–664 (2006)
9. Ruch, P., Baud, R.H., Geissbühler, A.: Learning-free text categorization. In: Dojat, M., Keravnou, E.T., Barahona, P. (eds.) AIME 2003. LNCS (LNAI), vol. 2780, pp. 199–208. Springer, Heidelberg (2003)
10. Ruch, P., Jimeno Yepes, A., Ehrler, F., Gobeill, J., Tbahriti, I.: Report on the trec 2006 experiment: Genomics track. In: TREC (2006)
11. 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. (ed.) *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA 1999)*, B.K., vol. 21(13-14), pp. 1193–1198 (2000)
12. Swain, M.J., Ballard, D.H.: Color indexing. *International Journal of Computer Vision* 7(1), 11–32 (1991)
13. Gass, T., Geissbuhler, A., Müller, H.: Learning a frequency-based weighting for medical image classification. In: *Medical Imaging and Medical Informatics (MIMI) 2007*, Beijing, China (2007)

An Interactive and Dynamic Fusion-Based Image Retrieval Approach by CINDI

M.M. Rahman¹, B.C. Desai¹, and P. Bhattacharya^{2,*}

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 image retrieval technique and the analysis of different runs of ImageCLEF 2007 submitted by the CINDI group. An interactive fusion-based search technique is investigated in both context and content-based feature spaces. For a context-based image search, keywords from associated annotation files are extracted and indexed based on the vector space model of information retrieval. For a content-based image search, various global and region-specific local image features are extracted to represent images at different levels of abstraction. Based on a user's relevance feedback information, multiple textual and visual query refinements are performed and weights are adjusted dynamically in a similarity fusion scheme. Finally, top ranked images are obtained by performing both sequential and simultaneous search processes in the multi-modal (context and content) feature space.

1 Introduction

CINDI research group has participated in the ad-hoc retrieval tasks of both general photographs and medical collections in ImageCLEF 2007 [1,2]. The effectiveness of interactive relevance feedback and dynamic fusion-based approaches are investigated in individual context and content-based feature spaces as well as in a combination of both feature spaces.

2 Context-Based Image Search

To incorporate context (associated annotation) based image search, indexing is performed by extracting keywords from associated annotation files of images and using the vector space model (VSM) of information retrieval (IR) [3]. In this model, the annotation part (document) D_j of images is represented as a vector in a N -dimensional space as $\mathbf{f}_{D_j} = [w_{j1} \ w_{j2} \ \cdots \ w_{jN}]^T$. The element w_{jk} denotes the *TF-IDF* weight [3] of a term $t_k \in \{t_1, t_2, \dots, t_N\}$ in the D_j . A query D_q is

* This work was partially supported by NSERC, IDEAS and Canada Research Chair grants.

also represented as a vector $\mathbf{f}_{D_q} = [w_{q1} \ w_{q2} \cdots \ w_{qN}]^T$. To compare D_q and D_j , the cosine similarity measure is applied as follows

$$\text{Sim}_{\text{text}}(D_q, D_j) = \cos(\mathbf{f}_{D_q}, \mathbf{f}_{D_j}) = \frac{\sum_{i=1}^N w_{qi} * w_{ji}}{\sqrt{\sum_{i=1}^N (w_{qi})^2} * \sqrt{\sum_{i=1}^N (w_{ji})^2}} \quad (1)$$

where, w_{iq} and w_{ij} are the weights of the term t_i in D_q and D_j respectively.

2.1 Textual Query Refinement

Multiple query reformulation based on different relevance feedback (RF) methods might provide different interpretations of a user's underlying information need. It has been demonstrated in [4,5] that if two query representation retrieve different sets of documents, significant improvement can be achieved by combining the retrieval result. Hence, we generate multiple query representations by applying various RF-based methods. For the first two methods, the well known *Rocchio* [6] and *Ide-dec-hi*[7] algorithms are used. These algorithms generally move a new query point toward relevant documents and away from irrelevant documents in a feature space [6]. After performing the query reformulation in a feedback iteration, the modified query vectors are termed here as $\mathbf{f}_{D_q}^m(\text{Rocchio})$ and $\mathbf{f}_{D_q}^m(\text{Ide})$. A simpler approach of query expansion is also considered based on the identification of top l useful terms from relevant annotation files. After identifying the most frequently occurred l terms, the query vector is updated as $\mathbf{f}_{D_q}^m(\text{Local1})$ by re-weighting its keywords based on the *TF-IDF* weighting scheme. The final query reformulation approach is based on expanding the original query with terms correlated to the query terms. Such correlated terms are those present in local clusters built from a local vocabulary T_l based on relevant documents [8,3]. For this, a correlation matrix $C_{(|T_l| \times |T_l|)}$ is constructed [9]. The element of this matrix $c_{u,v}$, is defined as

$$c_{u,v} = \frac{n_{u,v}}{n_u + n_v - n_{u,v}} \quad (2)$$

where, n_u and n_v are the number of local documents which contain terms t_u and t_v respectively, and $n_{u,v}$ is the number of local documents which contain both terms. If t_u and t_v have many co-occurrences in local documents, then the value of $c_{u,v}$ increases, and the documents are considered to be more correlated. Considering the u -th row in the matrix C (i.e., the row with all the correlations for a query term t_u), it returns the set of n largest correlation values $c_{u,l}$, where $l \neq u$. For a query D_q , we are normally interested in finding clusters only for the $|D_q|$ query terms. After extracting the additional terms for each query term, the query vector is updated as $\mathbf{f}_{D_q}^m(\text{Local2})$ by re-weighting its keywords based on the *TF-IDF* weighting scheme. In this work, $l = 5$ and $n = 3$ are utilized for the experiments.

3 Content-Based Image Search

For a content-based image search, various global, semi-global, region-specific local, and visual concept-based image features are extracted. It is assumed that the image features at different levels of abstraction are complementary in nature. For feature representation at a *global* level, the MPEG-7 based Edge Histogram Descriptor (EHD) and Color Layout Descriptor (CLD) [10] are utilized in this work. The EHD represents the spatial distribution of edges as a global shape feature and the CLD represents the spatial layout of images in a very compact form. Several moment-based color (e.g., mean and standard deviation of each color channel in *HSV*) and texture features (e.g., energy, maximum probability, entropy, contrast and inverse difference from grey-level co-occurrence matrices) are extracted from five overlapping sub-images and finally they are combined to form a joint *semi-global* feature vector. A region-based image retrieval approach is also considered in this work. Here, each image is automatically segmented by a K-means clustering technique into a set of homogeneous regions. The *local* region-specific color and texture related features of each image are utilized in an image level similarity matching function by considering individual region level similarity measures based on the Bhattacharyya distance [12]. Finally, images are also represented in a *visual concept*-based feature space based on image encoding from an automatically generated codebook of visual concepts [13]. In the present work, two-dimensional codebooks of size 400 (e.g., 20×20 units) are constructed from training image samples by applying a self-organizing map (SOM)-based clustering technique. The detailed representation of all the above features and their similarity matching functions are described in our previous works [11,12,13].

3.1 Visual Query Refinement

This section presents our visual query refinement approach based on RF that not only performs query point movement but also adjusts the distance matching functions. It is assumed that, all relevant images belong to a user's perceived semantic category and obey the Gaussian distribution to form a cluster in the feature space. The modified vector of a query image I_q at an iteration k is represented as the mean of the relevant image vectors as

$$\mathbf{f}_{I_q^x}^m(k) = \frac{1}{|R_k|} \sum_{\mathbf{f}_{I_l^x} \in R_k} \mathbf{f}_{I_l^x} \quad (3)$$

where, R_k be the set of relevant image vectors at iteration k and $x \in \{CLD, EHD, Moment, Concept\}$. The covariance matrices of the positive feature vectors are estimated as

$$\mathbf{C}_{(k)}^x = \frac{1}{|R_k| - 1} \sum_{l=1}^{|R_k|} (\mathbf{f}_{I_l^x} - \mathbf{f}_{I_q^x}^m(k)) (\mathbf{f}_{I_l^x} - \mathbf{f}_{I_q^x}^m(k))^T \quad (4)$$

After generating the mean vector and covariance matrix for a feature x , the individual Euclidean-based distance measure functions are adjusted with the following Mahalanobis distance measure [14] for query image I_q and database image I_j as

$$\text{Dis}_x(I_q, I_j) = (\mathbf{f}_{I_q^x}^m - \mathbf{f}_{I_j^x})^T \hat{\mathbf{C}}_{(k)}^{-1} (\mathbf{f}_{I_q^x}^m - \mathbf{f}_{I_j^x}) \quad (5)$$

The Mahalanobis distance differs from the Euclidean distance in that it takes into account the correlations of the feature attributes and is scale-invariant [14]. We did not perform any query refinement for region-specific features at this moment due to their variable feature dimension for variable number of regions in images.

The modified query vectors of both contextual and visual feature spaces are submitted to the system for the next iteration. In the following section, we propose a dynamically weighted linear combination of similarity fusion technique that updates both inter and intra modality feature weights for the next iteration to obtain a final ranked list of images either from an individual modality (e.g., text or image) or a combination of both in a single search.

4 Adaptive Fusion-Based Similarity Matching

For multi-modal retrieval purposes, let us consider q as a multi-modal query which has an image part as I_q and a document part as annotation file as D_q . In a linear combination scheme, the similarity between q and a multi-modal item j , which also has two parts (e.g., image I_j and text D_j), is defined as

$$\text{Sim}(q, j) = \omega_I \text{Sim}_I(I_q, I_j) + \omega_D \text{Sim}_D(D_q, D_j) \quad (6)$$

where ω_I and ω_D are inter-modality weights within the text or image feature space, which is subject to $0 \leq \omega_I, \omega_D \leq 1$ and $\omega_I + \omega_D = 1$. The image based similarity function is further defined as the linear combination of similarity measures as

$$\text{Sim}_I(I_q, I_j) = \sum_{IF} \omega_I^{IF} \text{Sim}_I^{IF}(I_q, I_j) \quad (7)$$

where $IF \in \{\text{global, semi - global, local, concept}\}$ and ω_I^{IF} are the weights within the different image representation schemes (e.g., intra-modality weights). On the other hand, the text based similarity is defined as the linear combination of similarity matching based on different query representation schemes.

$$\text{Sim}_D(D_q, D_j) = \sum_{QF} \omega_D^{QF} \text{Sim}_D^{QF}(D_q, D_j) \quad (8)$$

where $QF \in \{\text{Rocchio, Ide, Local1, Local2}\}$ and ω_D^{QF} are the weights within the different query representation schemes.

The effectiveness of the linear combination depends mainly on the choice of the different inter and intra-modality weights. Motivated by the data fusion and relevance feedback paradigms, we propose a dynamic weight updating method

by considering both precision and rank order information from top retrieved K images. In this approach, to update the inter-modality weights (e.g., ω_I and ω_D), the multi-modal similarity matching based on equation (6) with equal weights is performed first. After the initial retrieval, a user provides feedback about the relevant images from the top K returned images. For each ranked list based on individual similarity matching, the top K images are considered and the effectiveness of a feature (text or image) is estimated as

$$E(D \text{ or } I) = \frac{\sum_{i=1}^K \text{Rank}(i)}{K/2} * P(K) \quad (9)$$

where $\text{Rank}(i) = 0$ if image in the rank position i is not relevant based on user's feedback and $\text{Rank}(i) = (K - i)/(K - 1)$ for the relevant images. Hence, the function $\text{Rank}(i)$ monotonically decreases from one (if the image at rank position 1 is relevant) down to zero (e.g., for a relevant image at rank position K). On the other hand, $P(K) = R_K/K$ is the precision at top K , where R_k is the number of relevant images in the top K retrieved images. The raw performance scores obtained by the above procedure are then normalized by the total score as $\hat{E}(D) = \hat{\omega}_D = \frac{E(D)}{E(D)+E(I)}$ and $\hat{E}(I) = \hat{\omega}_I = \frac{E(I)}{E(D)+E(I)}$ to generate the updated text and image feature weights respectively. For the next iteration of retrieval with the same query, these modified weights are utilized for the multi-modal similarity matching function as

$$\text{Sim}(q, j) = \hat{\omega}_I \text{Sim}_I(I_q, I_j) + \hat{\omega}_D \text{Sim}_D(D_q, D_j) \quad (10)$$

This weight updating process might be continued as long as users provide relevant feedback information or until no changes are noticed due to the system convergence. In a similar fashion, the intra-modality weights (e.g., ω_D^{QF} and ω_I^{IF}) can be updated by considering the top K images in individual image only or text only result lists and following equation (9).

5 Simultaneous and Sequential Search Processes

Both simultaneous and sequential search approaches are investigated to obtain the final ranked list of retrieval results. For simultaneous approach, initially a multi-modal search is performed to rank the images based on equation (6) with equal inter modality weights. Next, a user's feedback about relevant and irrelevant images from the top retrieved K images is obtained for refining both textual and visual queries as described in Sections 2.1 and 3.1 and for dynamically updating the weights as described in Section 4. The modified textual and image query vectors are re-submitted to the system and the multi-modal similarity matching based on (10) is performed for the next iteration. The above process can be continued until the user is satisfied or the system converges.

Since a topic is represented with both keywords and visual features (e.g., example image), a search can be initiated either by using keywords or by using visual example images. We consider such a sequential search approach in which

combining the results of the text and image based retrieval is a matter of re-ranking or re-ordering the images in a text-based pre-filtered result set. In this approach, for a multi-modal query q with a document part as D_q , a textual search is performed at first to obtain a result set of images. Next, a user's feedback about relevant and irrelevant images are obtained from top retrieved K images and textual query refinement is performed based on different RF methods. The modified query vectors are re-submitted to the system and text-based similarity matching is performed based on equation (8) with equal weights. In consequent iterations, the equal weights are dynamically updated based on equation (9) for similarity matching. A visual only search based on equation (7) is performed on the top L images retrieved by the previous text-based search. Next, a user's feedback about relevant images from the top retrieved K images is provided to the system to perform visual query refinement and dynamic weight update as described in Sections 3.1 and 4. The search process can be continued for few iterations to obtain a final re-ranked image list. In all the cases, we used $K = 30$ and $L = 2000$ for the experimental purpose.

6 Result Analysis of the Submitted Runs

The retrieval results of different runs are shown in Table 1 and Table 2 for the ad-hoc retrieval of the photographic and medical collections respectively. In all these runs, only English was used as the source and target languages without any translation for the text-based retrieval approach. We submitted five different runs for the ad-hoc retrieval of the photographic collection, where the first two runs were based on text only search and the last three runs were based on multimodal search as shown in Table 1. For the first run *CINDI-TXT-ENG-PHOTO*, we performed only a manual text-based search without any query expansion as our base run. This run achieved a MAP score of 0.1529 and ranked 140th out of 476 different submissions (e.g., within the top 30%). Our second run *CINDI-TXT-QE-PHOTO* achieved the best MAP score (0.2637) among all our submitted runs and ranked 21st. In this run, we performed two iterations of manual feedback for textual query expansion and combination based on the dynamic weight updating scheme for the text only retrieval. The rest of the runs were based on multimodal search. For the third run *CINDI-TXT-QE-IMG-RF-RERANK*, we performed the sequential search approach with two iterations of manual feedback in both text and image-based searches. This run ranked 32nd in terms of a MAP score of 0.2336. For the fourth run *CINDI-TXTIMG-FUSION-PHOTO*, we performed a simultaneous retrieval approach without any feedback information with a linear combination of weights as $\omega_D = 0.7$ and $\omega_I = 0.3$ and two iterations of feedback were performed for the fifth run *CINDI-TXTIMG-RF-PHOTO*. Overall, the sequential search approach performed better compared to the simultaneous search. From the result of Table 1, we can observe that our adaptive multimodal runs did not improve the result when compared to the best run based on the textual query expansion approach. The main reason might be due to the fact that the system did not get enough feedback information for the

Table 1. Results of the General Photograph Retrieval task

Run ID	Modality QE/RF MAP BPREF			
CINDI-TXT-ENG-PHOTO	TXT	NOFB	0.1529	0.1426
CINDI-TXT-QE-PHOTO	TXT	FBQE	0.2637	0.2515
CINDI-TXT-QE-IMG-RF-RERANK	MIXED	FBQE	0.2336	0.2398
CINDI-TXTIMG-FUSION-PHOTO	MIXED	NOFB	0.1483	0.1620
CINDI-TXTIMG-RF-PHOTO	MIXED	FBQE	0.1363	0.1576

Table 2. Results of the Medical Image Retrieval task

Run ID	Modality QE/RF MAP R-prec			
CINDI-IMG-FUSION	IMAGE	NOFB	0.0333	0.0532
CINDI-IMG-FUSION-RF	IMAGE	FBQE	0.0372	0.0549
CINDI-TXT-IMAGE-LINEAR	MIXED	NOFB	0.1659	0.2196
CINDI-TXT-IMG-RF-LINEAR	MIXED	FBQE	0.0823	0.1168

function in (9) to perform effectively. In some cases, it might also affect negatively the final retrieval result when an image based search tries to complement a purely semantic oriented text-based search. We need to investigate these facts more in our future work.

Table 2 shows our results of the four different runs for the medical image retrieval task. A visual only search was performed based on the various image feature representation schemes without any feedback information for the first run *CINDI-IMG-FUSION*. For the second run *CINDI-IMG-FUSION-RF*, we performed only one iteration of manual feedback. For this run we achieved a MAP score of 0.0396, which is slightly better then the score (0.0333) achieved by the first run without any RF scheme. These two runs ranked among the top five results based on pure visual only run. For the third run *CINDI-TXT-IMAGE-LINEAR*, we performed a simultaneous retrieval approach without any feedback information with a linear combination of weights as $\omega_D = 0.7$ and $\omega_I = 0.3$ and for the fourth run *CINDI-TXT-IMG-RF-LINEAR*, two iterations of manual relevance feedback were performed similar to the last two runs of photographic retrieval task. From Table 2, it is clear that combining both modalities for the medical retrieval task is far better then using only a single modality (e.g., only image) in terms of the MAP score.

7 Conclusion

This paper examined the ad-hoc image retrieval approach of the CINDI research group for ImageCLEF 2007. We have submitted several runs with different combination of methods. In the future, we want to investigate more about multimodal fusion approaches and explore new ways to improve our current technique.

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 Photographic Retrieval Task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
2. Müller, H., Deselaers, T., Kim, E., Kalpathy, C., Jayashree, C., Thomas, D., William, H.: Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (2007)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Reading (1999)
4. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Proc. of the 2nd Text Retrieval Conference (TREC-2), vol. 500-215, pp. 243–252. NIST Special Publication (1994)
5. Lee, J.H.: Combining Multiple Evidence from Different Properties of Weighting Schemes. In: Proc. of the 18th Annual ACM-SIGIR, pp. 180–188 (1995)
6. Rocchio, R.: Relevance feedback in information retrieval. In: The SMART Retrieval System –Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall, Inc., Englewood Cliffs (1971)
7. Ide, E.: New experiments in relevance feedback. In: The SMART Retrieval System –Experiments in Automatic Document Processing, pp. 337–354. Prentice Hall, Inc., Englewood Cliffs (1971)
8. Attar, R., Fraenkel, A.: Local feedback in full-text retrieval systems. *Journal of the ACM (JACM)* 24 (3), 397–417 (1977)
9. Ogawa, Y., Morita, T., Kobayashi, K.: A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets and Systems* 39, 163–179 (1991)
10. Manjunath, B., Salembier, P., Sikora, T. (eds.): Introduction to MPEG-7 – Multimedia Content Description Interface, pp. 187–212. John Wiley Sons Ltd., Chichester (2002)
11. Rahman, M., Sood, V., Desai, B., Bhattacharya, P.: CINDI at ImageCLEF 2006: Image Retrieval & Annotation Tasks for the General Photographic and Medical Image Collections. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 715–724. Springer, Heidelberg (2007)
12. Rahman, M., Desai, B., Bhattacharya, P.: A Feature Level Fusion in Similarity Matching to Content-Based Image Retrieval. In: Proc. 9th International Conference on Information Fusion (2006)
13. Rahman, M., Desai, B., Bhattacharya, P.: Visual Keyword-based Image Retrieval using Correlation-Enhanced Latent Semantic Indexing, Similarity Matching & Query Expansion in Inverted Index. In: Proc. of the International Database Engineering & Applications Symposium (IDEAS 2006), pp. 201–208 (2006)
14. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, London (1990)

Using Pseudo-Relevance Feedback to Improve Image Retrieval Results

Mouna Torjmen, Karen Pinel-Sauvagnat, and Mohand Boughanem

IRIT, 118 Route Narbonne-31062 Toulouse Cedex 4 -France
{torjmen,sauvagna,bougha}@irit.fr

Abstract. In this paper, we propose a pseudo-relevance feedback method to deal with the photographic retrieval and medical retrieval tasks of ImageCLEF 2007. The aim of our participation to ImageCLEF is to evaluate a combination method using both english textual queries and image queries to answer to topics. The approach processes image queries and merges them with textual queries in order to improve results.

A first set of experiments using only textual information does not allow to obtain good results. To process image queries, we used the *FIRE* system to sort similar images using low level features, and we then used associated textual information of the top images to construct a new textual query. Results showed the interest of low level features to process image queries, as performance increased compared to textual queries processing.

Finally, best results were obtained combining the results lists of textual queries processing and image queries processing with a linear function.

1 Introduction

In Image Retrieval, one can distinguish two main approaches [1] : (1) Context Based Image Retrieval and (2) Content Based Image Retrieval:

- The context of an image is all information about the image coming from others sources than the image itself. For the time being, only textual information is used as context. The main problem of this approach is that documents can use different words to describe the same image or can use the same words to describe different concepts. Moreover image queries can't be processed.
- Content Based Image Retrieval (CBIR) systems use low-level image features to return images similar to an example image. The main problem of this approach is that visual similarity does not always correspond to semantic similarity (for example a CBIR system can return a picture of blue sky when the example image is a blue car).

Most of the image retrieval systems combine nowadays content and context retrieval, in order to take advantages of both methods. Indeed, it has been proved that combining text- and content-based methods for images retrieval always improves performance [2].

Images and textual information can be considered as independent and content and contextual information of queries can be combined in different ways:

- Image queries and textual queries can be processed separately and the two results lists are then merged using a linear function [3], [4].
- One can also use a pipeline approach: a first search is done using textual information or content information, and a filtering step is then processed using the other information type to exclude non-relevant images [5].
- Other methods use Latent Semantic Analysis (LSA) techniques to combine visual and textual information, but are not efficient [1] [6].

Some other works propose translation-based methods, in which content and context information are complementary. The main idea is to extract relations between images and text, and to use them to translate textual information to visual one and vice versa [7]:

- In [8], authors translate textual queries to visual ones.
- Authors of [9] propose to translate image queries to textual ones, and to process them using textual methods. Results are then merged with those obtained with textual queries. Authors in [10] also propose to expand the initial textual query by terms extracted thanks to an image query.

For the latter methods, the main problem to construct a new textual query or expand an initial textual query is term extraction. To do this, the main solution is *pseudo-relevance feedback*. Using pseudo-relevance feedback in context based image retrieval to process image queries is slightly different from classic pseudo-relevance feedback. The first step is to use a visual system to process image queries. Images obtained as results are considered as relevant and the associated textual information is then used to select terms in order to express a new textual query.

The work presented in this paper also proposes to combine context and content information to answer to the photographic retrieval and medical retrieval tasks. More precisely, we present a method to transform image queries to textual ones. We use *XFIRM* [11], a structured information retrieval system, to process english textual queries, and the *FIRE* system [12] to process image queries. Documents corresponding to the images returned by *FIRE* are used to extract terms that will form a new textual query.

The paper is organized as follows. In Section 2, we describe textual queries processing using the *XFIRM* system. In Section 3, we describe the image queries processing using in a first step, the *FIRE* system, and in a second step a pseudo-relevance feedback method. In Section 4, we present our combination method, which uses both results of the *XFIRM* and *FIRE* systems. Experiments and results for the two tasks (medical retrieval and photographic retrieval [13], [14]) are exposed in section 5. We discuss results in section 6 and finally we conclude in Section 7.

2 Textual Queries Processing

Textual information of collections used for the photographic and medical retrieval tasks [14] is organized using the XML language. In the indexing phase,

we decided to only use documents elements containing positive information: $\prec description \succ$, $\prec title \succ$, $\prec notes \succ$ and $\prec location \succ$.

We then used the *XFIRM* system [11] to process queries. *XFIRM* (*XML Flexible Information Retrieval Model*) uses a relevance propagation method to process textual queries in XML documents. Relevance values are first computed on *leaf nodes* (which contain textual information) and scores are then propagated along the document tree to evaluate *inner nodes* relevance values.

Let $q = t_1, \dots, t_n$ be a textual query composed of n terms. Relevance values of leaf nodes ln are computed thanks to a similarity function $RSV(q, ln)$.

$$RSV(q, ln) = \sum_{i=1}^n w_i^q * w_i^{ln}, \text{ where } w_i^q = tf_i^q \text{ and } w_i^{ln} = tf_i^{ln} * idf_i * ief_i \quad (1)$$

w_i^q and w_i^{ln} are the weights of term i in query q and leaf node ln respectively. tf_i^q and tf_i^{ln} are the frequency of i in q and ln , $idf_i = \log(|D|/(|di| + 1)) + 1$, with $|D|$ the total number of documents in the collection, and $|di|$ the number of documents containing i , and ief_i is the inverse element frequency of term i , i.e. $\log(|N|/|nf_i| + 1) + 1$, where $|nf_i|$ is the number of leaf nodes containing i and $|N|$ is the total number of leaf nodes in the collection.

idf_i allows to model the importance of term i in the collection of documents, while ief_i allows to model it in the collection of elements.

Each node n in the document tree is then assigned a relevance score r_n which is function of the relevance scores of the leaf nodes it contains and of the relevance value of the whole document.

$$r_n = \rho * |L_n^r|. \sum_{ln_k \in L_n} \alpha^{dist(n, ln_k) - 1} * RSV(q, ln_k) + (1 - \rho) * r_{root} \quad (2)$$

$dist(n, ln_k)$ is the distance between node n and leaf node ln_k in the document tree, i.e. the number of arcs that are necessary to join n and ln_k , and $\alpha \in]0..1]$ allows to adapt the importance of the $dist$ parameter. In all the experiments presented in the paper, α is set to 0.6.

L_n is the set of leaf nodes being descendant of n , and $|L_n^r|$ is the number of leaf nodes in L_n having a non-zero relevance value (according to equation 1). $\rho \in]0..1]$, inspired from work presented in [15], allows the introduction of document relevance in inner nodes relevance evaluation, and r_{root} is the relevance score of the *root* element, i.e. the relevance score of the whole document, evaluated with equation 2 with $\rho = 1$.

Finally, documents d_j containing relevant nodes are retrieved with the following relevance score:

$$r_{XFIRM}(d_j) = \max_{n \in d_j} r_n \quad (3)$$

Images associated to the documents are lastly returned by the system to answer to the retrieval tasks.

3 Image Queries Processing

To process image queries, we used a third-steps method: (1) a first step is to process images using the *FIRE* System [12], (2) we then use pseudo-relevance feedback to construct new textual queries, (3) the new textual queries are processed with the *XFIRM* system.

We first used the *FIRE* system to get the top K similar images to the image query. We then get the N associated textual documents (with $N \leq K$, because some images do not have associated textual information) and extracted the top L terms from them. To select the top L terms, we evaluated two formula to express the weight w_i of term t_i .

The first formula uses the frequency of term t_i in the N documents.

$$w_i = \sum_{j=1}^N t f_i^j \quad (4)$$

where $t f_i^j$ is the frequency of term t_i in document d_j .

The second formula uses terms frequency in the N selected documents, the number of documents in the N selected containing the term, and a normalized *idf* of the term in the whole collection.

$$w_i = [1 + \log(\sum_{j=1}^N t f_i^j)] * \frac{n_i}{N} * \frac{\log(\frac{D}{d_i})}{\log(D)} \quad (5)$$

where n_i is the number of documents in the N associated documents containing the term t_i , D is the number of all documents in the collection and d_i is the number of documents in the collection containing t_i .

The use of the $\frac{n_i}{N}$ parameter is based on the following assumption: a term occurring one time in n documents is more important and must be more relevant than a term occurring n times in one document. The *log* function is used on $\sum_{j=1}^N t f_i^j$ to emphasize the impact of the $\frac{n_i}{N}$ parameter.

We then construct a new textual query with the top L terms selected according to formula 4 or 5 and we process it using the *XFIRM* system (as explained in section 2).

In the photographic retrieval task, we obtained the following queries for topic Q48, with $K = 5$ and $L \leq 5$:

Textual query using equation 4: "south korea river"

Textual query using equation 5: "south korea night forklift australia"

The original textual query in english was: "vehicle in South Korea". As we can see, the query using equation 5 is more similar to the original query than the one using equation 4.

4 Combination Function

To evaluate the interest of using both content and context information, we combined results of image queries and textual queries processing and we evaluated

new relevance scores $r(d_j)$ for documents d_j :

$$r(d_j) = \lambda * (r_{XFIRM}(d_j)) + (1 - \lambda) * (r_{PRF}(d_j)) \quad (6)$$

where $r_{XFIRM}(d_j)$ is the relevance score of document d_j according to the *XFIRM* system (equation 3) and $r_{PRF}(d_j)$ is the relevance score of d_j according to the *XFIRM* system after image queries processing (see section 3).

In order to answer to both retrieval tasks, we then return all images associated to the top ranked documents. Figure 1 illustrates our approach.

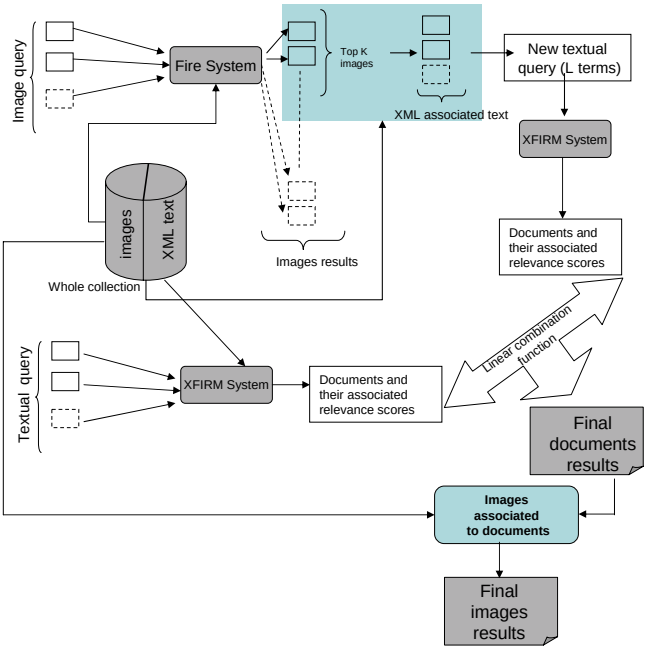


Fig. 1. Query processing with the combination of image and textual query processing approach

5 Evaluation and Results

5.1 Photographic Retrieval Task

– Evaluation of textual queries

We evaluated english textual queries using the *XFIRM* system with parameters $\rho = 0.9$ and $\rho = 1$. Results, which are almost the same, are presented in table 1.

– Evaluation of image queries

Table 2 shows results using the two formula described in section 3. We notice that the use of term frequency in selected documents is not enough, and that

the importance of the term in the collection need to be used in the term weighted function (results are better with equation 5 than with equation 4). If we now compare table 1 and table 2, we see that processing image queries with the *FIRE* system and our pseudo-relevance feedback system gives better results than using only the *XFIRM* system on textual queries. It shows the importance of visual features to retrieve images.

– **Combination of textual and image queries results**

Table 3 shows our results for the combination approach. For all these experiments, L is set to 5.

Let us first compare runs Runcomb1 and Runcomb4, which use eq. 4 and K=6, and eq. 5 and K=15. For both, we use $\rho = 1$ and $\lambda = 0.9$ for the combination. Results show that using eq. 5 with K=15 is more efficient than eq. 4 with K=6, which confirms results obtained using only image queries.

In order to evaluate the combination function, we then use eq. 5, and fix $\rho = 1$ and K=15. We test $\lambda = 0.5$ and $\lambda = 0.9$ (runs Runcomb3 and Runcomb4). Results are almost the same but combining equally the two sources of evidence gives slightly better results.

Finally, we vary $\rho = 0, 9$ and $\rho = 1$, and fix equation 5, $\lambda = 0.9$ in equation 6 and K=15 (runs Runcomb4 and Runcomb2). Better results are obtained with $\rho = 1$, which means that the document relevance should not be taken into account in the evaluation of inner nodes relevance values (equation 2).

Table 1. Textual queries results using the *XFIRM* system

Run-id	ρ	MAP	P10	P20	P30	Bpref	GMAP
RunText0609	0.9	0.0634	0.1400	0.1175	0.1133	0.0719	0.0039
RunText061	1	0.0633	0.1400	0.1175	0.1128	0.0719	0.0039

Table 2. Image queries results using pseudo-relevance feedback with the *FIRE* and *XFIRM* systems

Run-id	K	L	ρ	Eq.	MAP	P10	P20	P30	Bpref	GMAP
RunPRF061tf	6	5	1	eq. 4	0.063	0.140	0.117	0.113	0.071	0.003
RunPRF061tfnNidf	6	15	1	eq. 5	0.123	0.210	0.200	0.179	0.138	0.006
RunPRF0609tfnNidf	6	15	0.9	eq. 5	0.125	0.211	0.200	0.179	0.138	0.006

Table 3. Results using the combination function

Run-id	K	λ	ρ	Eq.	MAP	P10	P20	P30	Bpref	GMAP
RunComb1 6	6	0.9	1	eq. 4	0.103	0.150	0.124	0.118	0.091	0.031
RunComb2 6	6	0.9	0.9	eq. 5	0.109	0.143	0.129	0.126	0.096	0.029
RunComb3 15	15	0.5	1	eq. 5	0.135	0.221	0.198	0.183	0.140	0.035
RunComb4 15	15	0.9	1	eq. 5	0.130	0.210	0.198	0.186	0.145	0.026

5.2 Medical Retrieval Task

For this task, we only evaluated the combination method described in section 4. RComb09 uses equation 5 with $\rho = 1$, $K=15$, $L=10$ and $\lambda = 0.9$. RComb05, our official run, uses equation 4 with $\rho=1$, $K=6$, $L=5$ and $\lambda = 0.5$.

Results are significantly better for run RComb09. However, as many parameters are involved (K , L , λ and the equation used to select terms) it is difficult to conclude on which parameters impact the results. Further experiments are thus needed.

Table 4. Results of the Medical retrieval task

Run-id	Eq.	L	K	λ	MAP	R-prec	Bpref	P10	P30	P100	P500	P1000
RComb09	eq.5	10	15	0.9	0.110	0.141	0.213	0.166	0.152	0.144	0.067	0.041
RComb05	eq.4	5	6	0.5	0.048	0.070	0.168	0.05	0.075	0.058	0.058	0.038

6 Discussion

The number of textual information resources used to construct new textual queries from image queries (i.e the K number of images selected from FIRE results) has a great impact on results. Increasing K improves results by introducing relevant information. Another factor that impacts on results is the number of new query terms L . In our experiments, when K and L increase, the MAP metric also increases. Moreover, processing textual queries or images separately does not allow to obtain the best results: combining the two sources of evidence clearly improves results.

Finally, we'd like to conclude with the type of textual information used. In the Medical and Photographic Retrieval Tasks, textual information is encoded using the XML language, and as a consequence, we decided to use an XML-oriented information retrieval system to process textual queries (*XFIRM*). However, elements are not organized in a hierarchic way as in can be the case in XML documents (no ancestor-descendant relationships between nodes), and the functions used by the *XFIRM* system to evaluate nodes relevance may be not appropriate in that case. Other experiments are consequently needed with a plain-text information retrieval system. Combining the *XFIRM* system with the *FIRE* system may be however interesting with fully encoded-XML collections.

7 Conclusion and Future Work

We participated in the Photographic and Medical Retrieval Tasks of ImageCLEF 2007 in order to evaluate a method using a content- and context-based approach to answer to topics. We proposed a new pseudo-relevance feedback approach to process image queries and we tested an XML oriented system to process textual queries. Results showed the interest of combining the two sources of evidence (content and context) to answer to image retrieval.

In future work, we plan to:

- Add low level features results extracted from *FIRE* to the combination function in the Medical Retrieval Task, as visual features are very important in the medical domain.
- Sort images using concepts level features [16] instead of low level features to construct new textual queries in the Photographic Retrieval Task.
- Use specific domain ontology to expand textual queries (original textual queries and queries obtained with our pseudo-relevance feedback approach).

References

1. Westerveld, T.: Image retrieval: Content versus context. In: Content-Based Multimedia Information Access, RIAO 2000 Conference Proceedings, pp. 276–284 (2000)
2. Deselaers, T., Müller, H., Clogh, P., Ney, H., Lehmann, T.M.: The clef 2005 automatic medical image annotation task. *International Journal of Computer Vision* 74(1), 51–58 (2007)
3. Boll, S., Klas, W., Wandel, J.: A cross-media adaptation strategy for multimedia presentations. In: *ACM Multimedia* (1), pp. 37–46 (1999)
4. Jones, G.J.F., Burke, M., Judge, J., Khasin, A., Lam-Adesina, A.M., Wagner, J.: Dublin city university at clef 2004: Experiments in monolingual, bilingual and multilingual retrieval. In: *CLEF*, pp. 207–220 (2004)
5. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words (1999)
6. Zhao, R., Grosky, W.: Narrowing the semantic gap - improved text-based web document retrieval using visual features (2002)
7. Lin, W.C., Chang, Y.C., Chen, H.H.: Integrating textual and visual information for cross-language image retrieval: A trans-media dictionary approach. *Inf. Process. Manage.* 43(2), 488–502 (2007)
8. Lin, W.C., Chang, Y.C., Chen, H.H.: Integrating textual and visual information for cross-language image retrieval. In: *Proceedings of the Second Asia Information Retrieval Symposium*, pp. 454–466 (2005)
9. Chang, Y.C., Lin, W.C., Chen, H.H.: A corpus-based relevance feedback approach to 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. 592–601. Springer, Heidelberg (2006)
10. Maillot, N., Chevallet, J.P., Valea, V., Lim, J.H.: Ipal inter-media pseudo-relevance feedback approach to imageclef 2006 photo retrieval. In: *Working Notes for the CLEF 2006 Workshop*, 20–22 September, Alicante, Spain (2006)
11. Sauvagnat, K.: *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. PhD thesis, Toulouse: Paul Sabatier University (2005)
12. Deselaers, T., Keysers, D., Ney, H.: FIRE — flexible image retrieval engine: ImageCLEF 2004 evaluation. In: *CLEF Workshop* (2004) (2004)
13. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary (2007)
14. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary (2007)

15. Mass, Y., Mandelbrod, M.: Experimenting various user models for XML retrieval. In: [17] (2005)
16. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: MULTIMEDIA 2006: Proceedings of the 14th annual ACM international conference on Multimedia, pp. 421–430. ACM Press, New York (2006)
17. Fuhr, N., Lalmas, M., Malik, S., Kazai, G.: INEX 2005 workshop proceedings (2005)

Overview of the CLEF-2007 Cross-Language Speech Retrieval Track

Pavel Pecina¹, Petra Hoffmannová¹, Gareth J.F. Jones², Ying Zhang²,
and Douglas W. Oard³

¹ MFF UK, Malostranske namesti 25, Room 422
Charles University, 118 00 Praha 1, Czech Republic
pecina@ufal.mff.cuni.cz, hoffmannova@knih.mff.cuni.cz

² School of Computing
Dublin City University, Dublin 9, Ireland
gjones@computing.dcu.ie, yzhang@computing.dcu.ie

³ College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
oard@umd.edu

Abstract. The CLEF-2007 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. Six teams participated in the English evaluation, performing both monolingual and cross-language searches of ASR transcripts, automatically generated metadata, and manually generated metadata. Four teams participated in the Czech evaluation, performing monolingual searches of automatic speech recognition transcripts.

1 Introduction

The 2007 Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) track was the third and final year for evaluation of ranked retrieval from spontaneous conversational speech from an oral history collection at CLEF. As in the CLEF 2006 CL-SR task [1], automatically transcribed interviews conducted in English could be searched using queries in one of six languages, and automatically transcribed interviews conducted in Czech could be searched using queries in one of two languages. New relevance judgments for additional topics were created to expand the Czech collection in 2007. The English collection used in 2007 was the same as that used in 2006. As in CLEF 2005 and CLEF 2006, the English task was based on a known-boundary condition for topically coherent segments. The Czech task was based on a unknown-boundary condition in which participants were required to identify a time stamp for the beginning of each distinct topically relevant passage.

The remainder of this paper is organized as follows. Section 2 describes the English task and summarizes the results for the submitted runs. Section 3 does

the same for the Czech task. The paper concludes in Section 4 with a brief recap of what has been learned across all three years of the CLEF CL-SR track.

2 English Task

The structure of the CLEF 2007 CL-SR English task was identical to that used in 2006, which we review here briefly (see [1] for more details).

2.1 Segments

The “documents” searched in the English task are 8,104 segments that were designated by professional indexers as topically coherent. A detailed description of the structure and fields of the English segment collection is given in the 2005 track overview paper [2]. Automatically generated transcripts from two Automatic Speech Recognition (ASR) systems are available. The ASRTEXT2006B 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. Only 7,378 segments have text in this field. For the remaining 726 segments, no ASR output was available from that system, so in those cases the ASRTEXT2006B field includes content identical to the ASRTEXT2004A field which was generated using an earlier less accurate transcription system (with a 35% mean word error rate). An extensive set of manually and automatically generated metadata is also available for each segment.

2.2 Topics

The same 63 training topics and 33 evaluation topics were used for the English task this year as had been used in 2006. Participating teams were asked not to use the evaluation topics for system tuning. Translations into Czech, Dutch, French, German, and Spanish had been created by native speakers of those languages. Participating teams were asked to submit runs for 105 topics (the 63 training topics, the 33 evaluation topics, and 9 other topics), but results are reported here only for the 33 evaluation topics.

2.3 Evaluation Measure

As in the CLEF-2006 CL-SR track, we report uninterpolated Mean Average Precision (MAP) as the principal measure of retrieval effectiveness. Version 8.0 of the `trec_eval` program was used to compute this measure.¹ The Wilcoxon signed-rank signed test was employed for evaluation of significance.

2.4 Relevance Judgments

We reused the relevance judgments from the English task of CLEF-2005, which had been created from multi-scale and multi-level relevance assessments performed by subject matter experts [2]. These judgments were conflated into binary

¹ The `trec_eval` program is available from http://trec.nist.gov/trec_eval/.

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. The resulting binary relevance judgments were filtered to remove segments which had been judged but had not been included in the test collection. This resulted in a total of 20,560 binary judgments across the 33 topics, among which 2,449 (12%) are relevant.²

2.5 Techniques

This section gives a brief description of the methods used by each team participating in the English task. Additional details are available in each team's paper.

Brown University (BLLIP). The Brown Laboratory for Linguistic Information Processing (BLLIP) team extended the basic Dirichlet-smoothed unigram IR model to incorporate bigram mixing and collection smoothing. In their enhanced language model, the bigram and unigram models were mixed using a tunable mixture weight over all documents. They attempted linearly mixing the test collection with two larger text corpora, 40,000 sentences from the Wall Street Journal and 450,000 sentences from the North American News Corpus, in order to alleviate the sparse data problems in the case of small collections. They observed that bigram statistics appeared to have greater impact with pseudo-relevance feedback than without. The collection smoothing approach clearly provided a substantial improvement.

Dublin City University (DCU). Dublin City University concentrated on the issues of topic translation, combining this with search field combination and pseudo-relevance feedback methods used for their CLEF 2006 submissions. Non-English topics were translated into English using the Yahoo! BabelFish free online translation service and with domain-specific translation lexicons gathered automatically from Wikipedia. Combination of multiple fields using the BM25F variant of Okapi weights was explored. Additionally, the DCU team integrated their information retrieval methods based on the Okapi model with summary-based pseudo-relevance feedback.

University of Amsterdam (UVA). The University of Amsterdam explored the use of character n -gram tokenization to improve the retrieval of documents using automatically generated text, as well as the combination of manually generated with automatically generated text. They reported that $n = 4$ provided the

² For CLEF-2006, a less careful filtering resulted in 28,223 binary judgments, of which 2,450 were relevant. The only difference in the relevant subset is that the 2007 judgments contain 33 rather than 34 relevant for topic 3032. Since the computation of uninterpolated MAP by `trec_eval` is affected only by the relevant subset, uninterpolated MAP values from 2006 and 2007 can reasonably be directly compared without adjustment for differences in the relevance judgments.

best retrieval effectiveness when a cross-word overlapping n -gram tokenization strategy was used. The field combination was done using the Indri query language, in which varying weights were assigned to different fields. Cross-language experiments were conducted using Dutch topics that were automatically translated into English using two different online tools, SYSTRAN and FreeTranslation. The translations generated from each MT system were then combined as a ‘bag-of-words’ English query.

University of Chicago (UC). The University of Chicago team focused on the contribution of automatically assigned thesaurus terms to retrieval effectiveness and the utility of different query translation strategies. For French–English cross-language retrieval, they adopted two query translation strategies: MT-based translation using the publicly available translation tool provided by Google, and dictionary-based translation. Their dictionary-based translation procedure applied a backoff stemming strategy in order to support matching with highest precision between the query terms and the bilingual word list. They noted that 27% of the French query terms remained untranslated and were thus retained.

University of Jaén (SINAI). The SINAI group at the University of Jaén investigated the effect of selection of different fields on retrieval effectiveness. An information gain measure was employed to select the best XML tags in the document collection. The tags with higher information gain values were selected to compose the final collection. Their experiments were conducted with the Lemur retrieval information system using applying KL divergence. French, German, Spanish and Dutch topics were translated to English using a translation module, SINTRAM, which works with different online machine translation systems and combines the different translations based on heuristics.

University of Ottawa (UO). The University of Ottawa used weighted summation of normalized similarity measures to combine 15 different weighting schemes from two IR systems (Terrier and SMART). Two query expansion techniques, one based on the thesaurus and the other one on blind relevance feedback, were examined. In their cross-language experiments, the queries were automatically translated from French and Spanish into English by combining the results of multiple online machine translation tools. Results for an extensive set of locally scored runs were also reported.

2.6 Results

Table 1 summarizes the evaluation results for all 29 official runs averaged over the 33 evaluation topics, listed in descending order of MAP. These 29 runs were further categorized into four groups based on the query language used (English or non-English) and the document fields (automatic-only or at least one manual assigned) indexed: 9 automatic-only monolingual runs, 6 automatic-only cross-language runs, 9 monolingual runs with manually assigned metadata, and 5 cross-language runs with manually assigned metadata.

Table 1. Evaluation results for all English official runs. MK = MANUALKEYWORD (Manual metadata), SUM = SUMMARY (Manual metadata), AK1 = AUTOKEYWORD2004A1 (Automatic), AK2 = AUTOKEYWORD2004A2, ASR03 = ASRTEXT2003A (Automatic), ASR04 = ASRTEXT2004A (Automatic), ASR06A = ASRTEXT2006A (Automatic), ASR06B = ASRTEXT2006B (Automatic), and ALL = all fields

Run ID	MAP	Lng	Query	Document	Fields	Site
dcuEnTDNmanual	0.2847	EN	TDN	MK,SUM		DCU
uoEnTDtManF1	0.2761	EN	TD	MK,SUM		UO
brown.TDN.man	0.2577	EN	TDN	MK,SUM		BLLIP
dcuEnTDmanualauto	0.2459	EN	TD	MK,SUM,ASR06B		DCU
brown.TD.man	0.2366	EN	TD	MK,SUM		BLLIP
brown.T.man	0.2348	EN	T	MK,SUM		BLLIP
UvA_4.enopt	0.2088	EN	TD	MK,SUM,ASR06B		UVA
dcuFrTDmanualauto	0.1980	FR	TD	MK,SUM,ASR06B		DCU
UvA_5.nlopt	0.1408	NL	TD	MK,SUM,AK2,ASR06B		UVA
uoEnTDtQExF1	0.0855	EN	TD	AK1,AK2,ASR04		UO
uoEnTDtQExF2	0.0841	EN	TD	AK1,AK2,ASR04		UO
brown.TDN.auto	0.0831	EN	TDN	AK1,AK2,ASR06B		BLLIP
dcuEnTDauto	0.0787	EN	TD	AK1,AK2,ASR06B		DCU
brown.TD.auto	0.0785	EN	TD	AK1,AK2,ASR06B		BLLIP
SinaiSp100	0.0737	ES	TD	ALL		SINAI
dcuFrTDauto	0.0636	FR	TD	AK1,AK2,ASR06B		DCU
uoEsTDtF2	0.0619	ES	TD	AK1,AK2,ASR04		UO
uoFrTDtF2	0.0603	FR	TD	AK1,AK2,ASR04		UO
SinaiFr100	0.0597	FR	TD	ALL		SINAI
SinaiEn100	0.0597	EN	TD	ALL		SINAI
SinaiSp050	0.0579	ES	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B		SINAI
UCkwENTD	0.0571	EN	TD	AK1,AK2,ASR06B		UC
SinaiEn050	0.0515	EN	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B		SINAI
UCbaseENTD1	0.0512	EN	TD	ASR06B		UC
UvA_2.en4g	0.0444	EN	TD	AK2,ASR06B		UVA
UvA_1.base	0.0430	EN	TD	ASR06B		UVA
UCkwFRTD1	0.0406	FR	TD	AK1,AK2,ASR06B		UC
UvA_3.nl4g	0.0400	NL	TD	AK2,ASR06B		UVA
UCbaseFRTD1	0.0322	FR	TD	ASR06B		UC

Automatic-Only Monolingual Runs. Teams were required to run at least one monolingual condition using the title (T) and description (D) fields of the topics and indexing only automatically generated fields; the best of these “required runs” for each team are shown in bold in Tables 1 and 2 to facilitate comparison of results between different teams. The University of Ottawa (0.0855), Dublin City University (0.0787), and the BLLIP team (0.0785) reported comparable results (no significant difference at the 95% confidence level). These results are statistically significant better than those reported by the next two teams, the University of Chicago (0.0571) and the University of Amsterdam (0.0444), which were statistically indistinguishable from each other.

Table 2. Evaluation results for automatic English monolingual runs. Bold runs are the required condition. AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR03 = ASRTEXT2003A, ASR04 = ASRTEXT2004A, ASR06A = ASRTEXT2006A, and ASR06B = ASRTEXT2006B.

Run ID	MAP	Lng	Query	Document Fields	Site
uoEnTDtQExF1	0.0855	EN	TD	AK1,AK2,ASR04	UO
uoEnTDtQExF2	0.0841	EN	TD	AK1,AK2,ASR04	UO
brown.TDN.auto	0.0831	EN	TDN	AK1,AK2,ASR06B	BLLIP
dcuEnTDauto	0.0787	EN	TD	AK1,AK2,ASR06B	DCU
brown.TD.auto	0.0785	EN	TD	AK1,AK2,ASR06B	BLLIP
UCkwENTD	0.0571	EN	TD	AK1,AK2,ASR06B	UC
UCbaseENTD1	0.0512	EN	TD	ASR06B	UC
UvA_2_en4g	0.0444	EN	TD	AK2,ASR06B	UVA
UvA_1_base	0.0430	EN	TD	ASR06B	UVA

Table 3. Evaluation results for automatic cross-language runs. AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR04 = ASRTEXT2004A, and ASR06B = ASRTEXT2006B.

Run ID	MAP	Lng	Query	Document Fields	Site
dcuFrTDauto	0.0636	FR	TD	AK1,AK2,ASR06B	DCU
uoEsTDtF2	0.0619	ES	TD	AK1,AK2,ASR04	UO
uoFrTDtF2	0.0603	FR	TD	AK1,AK2,ASR04	UO
UCkwFRTD1	0.0406	FR	TD	AK1,AK2,ASR06B	UC
UvA_3_nl4g	0.0400	NL	TD	AK2,ASR06B	UVA
UCbaseFRTD1	0.0322	FR	TD	ASR06B	UC

Automatic-Only Cross-Language Runs. As shown in Table 3, the best result (0.0636) for cross-language runs on automatically generated indexing data (a French–English run from Dublin City University) achieved 81% of the monolingual retrieval effectiveness with comparable conditions (0.0787 as shown in Table 2).

Monolingual Runs with Manual Metadata. For monolingual TD runs on manually generated indexing data, the University of Ottawa achieved the best result (0.2761), which is statistically significantly better than all other runs under comparable conditions, as shown in Table 4. For TDN runs, the DCU result (0.2847) is not statistically significantly better than that obtained by BLLIP (0.2577).

Cross-Language Runs with Manual Metadata. The evaluation results for cross-language runs on manually generated indexing data are shown in Table 5. The best cross-language result (0.1980), representing 81% of monolingual retrieval

Table 4. Evaluation results for monolingual English runs with manual metadata. MK = MANUALKEYWORD, SUM = SUMMARY, AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR04 = ASRTEXT2004A, ASR06A = ASRTEXT2006A, ASR06B = ASRTEXT2006B, and ALL = all fields.

Run ID	MAP	Lng	Query	Document	Fields	Site
dcuEnTDNmanual	0.2847	EN	TDN	MK,SUM		DCU
uoEnTDtManF1	0.2761	EN	TD	MK,SUM		UO
brown.TDN.man	0.2577	EN	TDN	MK,SUM		BLLIP
dcuEnTDmanualauto	0.2459	EN	TD	MK,SUM,ASR06B		DCU
brown.TD.man	0.2366	EN	TD	MK,SUM		BLLIP
brown.T.man	0.2348	EN	T	MK,SUM		BLLIP
UvA_4_enopt	0.2088	EN	TD	MK,SUM,ASR06B		UVA
SinaiEn100	0.0597	EN	TD	ALL		SINAI
SinaiEn050	0.0515	EN	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B		SINAI

Table 5. Evaluation results for cross-language runs with manual metadata. MK = MANUALKEYWORD, SUM = SUMMARY, AK1 = AUTOKEYWORD2004A1, AK2 = AUTOKEYWORD2004A2, ASR04 = ASRTEXT2004A, ASR06A = ASRTEXT2006A, ASR06B = ASRTEXT2006B, and ALL = all fields.

Run ID	MAP	Lng	Query	Document	Fields	Site
dcuFrTDmanualauto	0.1980	FR	TD	MK,SUM,ASR06B		DCU
UvA_5_nlopt	0.1408	NL	TD	MK,SUM,AK2,ASR06B		UVA
SinaiSp100	0.0737	ES	TD	ALL		SINAI
SinaiFr100	0.0597	FR	TD	ALL		SINAI
SinaiSp050	0.0579	ES	TD	SUM,AK1,AK2,ASR04,ASR06A,ASR06B		SINAI

effectiveness under comparable conditions (0.2459 shown in Table 4), was achieved by DCU’s French-English run.

3 Czech Task

The structure of the Czech task was quite similar to the one used in the 2006, with differences which we describe in the following subsections. Further details can be found in the 2006 track overview paper [1].

3.1 Interviews

A “quickstart” collection was generated from the same set of 357 Czech interviews as in 2006. It contained 11,377 overlapping passages with the following fields:

DOCNO. Containing a unique document number in the same format as the start times that systems were required to produce in a ranked list.

INTERVIEWDATA. Containing 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.

ASRSYSTEM. Specifying the type of the ASR transcript, where “2004” and “2006” denote colloquial and formal Czech transcripts respectively.

CHANNEL. Specifying which recorded channel (left or right) was used to produce the transcript.

ASRTEXT. Containing words in order from the transcript selected by ASRSYSTEM and CHANNEL for a passage beginning at the start time indicated in DOCNO.

The average passage duration in the default 2007 quickstart collection is 3.75 minutes, and each passage has a 33% overlap with the subsequent passage (i.e., passages begin about every 2.5 minutes).

No thesaurus terms (neither manual nor automatic, neither English nor Czech) were distributed with the collection this year because it was not practical to correct the time misalignment that was present in the 2006 quickstart collection for the manually assigned thesaurus terms (and because the available automatically assigned thesaurus terms had not proven to be useful in 2006).

3.2 Topics

A total of 29 training topics and 42 evaluation topics were selected as follows. Participating teams were asked to submit results for a total of 118 topics: 105 topics from 2006 that had originally been created for the English collection, 10 topics from 2006 that were variants of 10 of the English topics that were “broadened” in a way that we expected to result in more matches in the Czech collection, and 3 new broadened topics that were constructed this year. For example, topic 1187 (Title: “IG Farben Labor Camps”) was broadened to create topic 4003 (Title: “Labor Camps”). All of these topics were originally created in English and then translated into Czech by native speakers.³ Some minor errors in the Czech translations from last year were corrected.⁴ No teams used the English topics this year; all official runs with the Czech collection were monolingual.

Two of the 118 topics were used for assessor training and excluded from the evaluation, 29 topics were available for training systems (with relevance judgments from 2006), and 50 of the remaining 87 topics were initially selected as possible evaluation topics. This set of 50 includes all available topics that were not used for assessor or system training for which at least 6 relevant passages were identified during the search-guided assessment phase. This cutoff at six segments was selected to balance quantization noise in the evaluation measure with the risk of

³ Dutch, French, German and Spanish versions are also available for the topics that were designed originally for the English task, but the 13 broadened topics have not been translated into those languages.

⁴ The corrected topics are 1259, 1282, 1551, 14313, and 24313. Of these, only topic 14313 had been selected as an evaluation topic in the 2006 Czech task. None of these have been used as evaluation topics in any year of the English task.

sampling error that would result from too few topics. An additional “pooled” assessment process was conducted after submission of results by participating teams to judge highly-ranked passages for which judgments had not been recorded during search-guided assessment. This pooled assessment process was completed for 42 of the 50 topics in the available time, so those 42 were chosen as the evaluation topics for the 2007 Czech task.

3.3 Evaluation Measure

The evaluation measure used for the Czech task was the same as in 2006: mean Generalized Average Precision (mGAP). This measure was originally designed to accommodate human assessments of partial relevance [3]. In our case, the human assessments are binary but the degree of match to those assessments can be partial. An exact match between the system-specified start time and the closest assessor-assigned start time yielded full credit for the match, with a linear decay to zero credit for system start time errors of plus or minus 90 seconds from the nearest assessor-assigned start time.⁵ The Wilcoxon signed-rank signed test was employed for evaluation of significance.

3.4 Relevance Judgments

Relevance judgments were completed at Charles University in Prague for the 42 evaluation topics this year under the same conditions as in 2006 by the same six relevance assessors. A total of 2,389 start (and end) times for relevant passages were identified, thus yielding an average of 56 relevant passages per topic (minimum 6, maximum 199). Table 6 shows the number of relevant start times for each of the 42 evaluation topics. A total of 34 of these 42 topics are also present in the CLEF CL-SR English task collection (as training, evaluation, or unused topics; the exceptions are 8 broadened topics, which are the 4000-series).

3.5 Techniques

All participating teams employed existing information retrieval systems to perform monolingual retrieval and submitted total of 15 runs for official scoring. To facilitate cross-team comparisons, each participating team submitted at least one run with the quickstart collection and with queries that were automatically created from the title and description topic fields. The narrative topic field was used only by University of West Bohemia. Most teams used only automatically generated queries; manual query construction was performed only by Charles University. The University of West Bohemia also used the quickstart scripts with different parameters to generate another collection for some experiments.

⁵ The window size was incorrectly reported as plus or minute 150 seconds in the 2006 CL-SR track overview paper, but a 90-second window was actually used in both 2006 and 2007.

Table 6. Number of relevant passages identified for each of the evaluation topics

Topic #	rel	Topic #	rel	Topic #	rel	Topic #	rel
1192	18	2265	113	3019	14	4005	68
1345	12	2358	126	3021	16	4006	135
1554	46	2384	37	3022	29	4007	51
1829	6	2404	8	3023	78	4009	10
1897	31	3000	41	3024	105	4011	132
1979	17	3001	102	3026	33	4012	61
2000	114	3002	95	3027	86	14313	17
2006	63	3007	107	3028	199	15601	108
2012	90	3008	53	3032	9	15602	25
2185	25	3010	18	4001	35		
2224	63	3016	40	4004	13		

Brown University (BLLIP). The Brown University system was based on a language model paradigm and implemented using Indri. A unigram language model, Czech-specific stemming, and pseudo-relevance feedback were applied in three officially submitted runs.

Charles University (CUNI). The Charles University team performed experiments with Indri using blind relevance feedback, stopword removal, and lemmatization obtained using a morphological analysis system that also performed part-of-speech tagging. The team submitted four official runs; two of which employed manual query construction.

University of Chicago (UC). The University of Chicago employed the InQuery information retrieval system with stopword removal and three different stemming approaches: no stemming, light stemming, and aggressive stemming. Three runs were submitted for official scoring.

University of West Bohemia (UWB). The University of West Bohemia employed a TF*IDF model implemented in Lemur with blind relevance feedback. Five runs were submitted for official scoring which differed in methods used for word normalization (none, lemmatization, stemming), in formulas used for term weighting (Raw TF, BM25), and in the topic fields used (TDN, TD).

Results. A computation error was discovered in the mGAP scoring script that was corrected after the CLEF-2007 meeting. Corrected results for all official runs (evaluated on 42 topics) are reported in Table 7, with bold indicating the highest-scoring run by each team with standard conditions (TD queries, standard quick-start collection),⁶ and the Charles University and University of West Bohemia

⁶ Corrected scores generally improved slightly, and the only reversal in system preference order was between two systems separated by 0.0001 in both the original and the corrected scores.

Table 7. Corrected scores for Czech official runs (Query language: CZ, Document fields: ASR2006, 90-second window)

Run name	mGAP score	Query construction	Topic fields	Term normalization	Site name
UWB_2-1.tdn.l	0.0274	Auto	TDN	lemma	UWB
UWB_3-1.tdn.l	0.0241	Auto	TDN	lemma	UWB
UWB_2-1.td.s	0.0229	Auto	TD	stem	UWB
UCcsaTD2	0.0213	Auto	TD	aggressive stem	UC
UCcslTD1	0.0196	Auto	TD	light stem	UC
prague04	0.0195	Auto	TD	lemma	CUNI
prague01	0.0192	Auto	TD	lemma	CUNI
prague02	0.0183	Manual	TD	lemma	CUNI
UWB_3-1.td.l	0.0134	Auto	TD	lemma	UWB
UWB_2-1.td.w	0.0132	Auto	TD	none	UWB
UCunstTD3	0.0126	Auto	TD	none	UC
brown.s.f	0.0113	Auto	TD	light stem	BLLIP
brown.sA.f	0.0106	Auto	TD	aggressive stem	BLLIP
prague03	0.0098	Manual	TD	none	CUNI
brown.f	0.0049	Auto	TD	none	BLLIP

papers in this volume report corrected mGAP scores as well. The effect of term normalization handling the rich Czech morphology is quite significant. The runs employing any type of term normalization (stemming or lemmatization) outperform systems indexing only original word forms with no normalization by 69–131%. The scores of directly comparable runs are given in Table 8, all the differences are statistically significant at a 95% confidence level.

Three quantization factors are present in the Czech evaluation: (1) the 15-second resolution of assessor-assigned start times; (2) the 90-second window size for mGAP computation, and (3) the 150-second spacing between passage start times in the standard quickstart collection. The 150-second passage start time spacing is clearly somewhat problematic when coupled with a 90-second evaluation window size. The University of West Bohemia demonstrated the effect by reducing the passage start time spacing to 75 seconds (the UWB_2-1 runs, in which the average passage duration was also reduced to 2.5 minutes). This yielded an apparent 14% increase in mGAP (compare UWB_2-1.tdn.l: mGAP=0.0274 and UWB_3-1.tdn.l: mGAP=0.0241) that turned out not to be statistically significant (perhaps because of quantization noise).

Although we compute evaluation results only from start times, our assessors marked both start and end times. The average duration of assessor-marked relevant passages is 2.83 minutes, which seems to be somewhat better matched to the 2.5 minutes passages used in the University of West Bohemia’s alternate condition (2.5 minutes for UWB_2-1.tdn.l vs. 3.75 minutes for UWB_3-1.tdn.l and all runs from other sites).

The Charles University team reported on the first experiments with interactive use of the Czech collection. Their best run based on manual query construction

Table 8. Comparison of systems with and without term normalization (Topic fields: TD, corrected results)

Run name	mGAP score	mGAP increase	Query construction	Term normalization	Site name
UWB.2-1_td_s	0.0229	+73%	Auto	stem	UWB
UWB.2-1_td_w	0.0132		Auto	none	UWB
UCcsaTD2	0.0213	+69%	Auto	aggressive stem	UC
UCunstTD3	0.0126		Auto	none	UC
prague02	0.0183	+87%	Manual	lemma	CUNI
prague03	0.0098		Manual	none	CUNI
brown.s.f	0.0113	+131%	Auto	light stem	BLLIP
brown.f	0.0049		Auto	none	BLLIP

(prague02) turned out to be statistically indistinguishable from a run under comparable conditions from the same team with queries that were generated automatically (prague04).

4 Conclusion and Future Plans

Like all CLEF tracks, the CL-SR track had three key goals: (1) to develop evaluation methods and reusable evaluation resources for an important information access problem in which cross-language access is a natural part of the task, (2) to generate results that can provide a strong baseline against which future research results with the same evaluation resources can be compared, and (3) to foster the development of a research community with the experience and expertise to make those future advances. In the case of the CL-SR track, those goals have now been achieved. Over three years, research teams from 14 universities in 6 countries submitted 123 runs for official scoring, and many additional locally scored runs have been reported in papers published by those research teams. The resulting English and Czech collections are the first standard information retrieval test collections for spontaneous conversational speech, unique characteristics of the English collection have fostered new research comparing searches based on automatic speech recognition and manually assigned metadata, and unique characteristics of the Czech collection have inspired new research on evaluation of information retrieval from unsegmented speech.

Now that the CL-SR track has been completed, these new CLEF test collections will be made available to nonparticipants through the Evaluations and Language Resources Distribution Agency (ELDA). The training data for the automatic speech retrieval systems that were used to generate the transcripts in those collections is also expected to become available soon, most likely through the Linguistic Data Consortium (LDC). It is our hope that these resources will be used together to investigate more closely coupled techniques than have been possible to date with just the present CLEF CL-SR test collections. Looking further forward, we believe that it is now time for the information retrieval research

community to look beyond oral history to other instances of spontaneous conversational speech such as recordings of meetings, historically significant telephone conversations, and broadcast conversations (e.g., radio “talk shows”). We also believe that it would be productive to begin to explore application of some of the technology developed for this track to improve access to a broad range of oral history collections and similar cultural heritage materials (e.g., interviews contained in broadcast archives). Together, these directions for future work will likely continue to extend the legacy and impact of this initial investment in exploring the retrieval of information from spontaneous conversational speech.

Acknowledgments

This year’s 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 Prague without whom the Czech collection simply would not exist, to Scott Olson for helping to prepare the English collection this year, to Ayelet Goldin and Jianqiang Wang for their timely help with critical details of the Czech relevance assessment and scoring process, to Pavel Ceske for creating the new Czech scoring script, to Jan Hajic for his support and advice throughout, and to Carol Peters for her seemingly endless patience. This work has been supported in part by NSF IIS award 0122466 (MALACH), by the Ministry of Education of the Czech Republic, projects MSM 0021620838 and #1P05ME786, and by the European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD—project MultiMATCH contract IST-033104. The authors are solely responsible for the content of this paper.

References

1. Oard, D.W., Wang, 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
2. 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)
3. Kekalainen, J., Jarvelin, K.: Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology* (2002)

A Dirichlet-Smoothed Bigram Model for Retrieving Spontaneous Speech

Matthew Lease and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI USA
{mlease,ec}@cs.brown.edu

Abstract. We present two simple but effective smoothing techniques for the standard language model (LM) approach to information retrieval [12]. First, we extend the unigram Dirichlet smoothing technique popular in IR [17] to bigram modeling [16]. Second, we propose a method of *collection expansion* for more robust estimation of the LM prior, particularly intended for sparse collections. Retrieval experiments on the MALACH archive [9] of automatically transcribed and manually summarized spontaneous speech interviews demonstrates strong overall system performance and the relative contribution of our extensions¹.

1 Introduction

In the language model (LM) paradigm for information retrieval (IR), a document's relevance is estimated as the probability of observing the query string as a random sample from the document's underlying LM [12]. The standard unigram LM approach has been shown to have a strong theoretical connection to TF-IDF [17] and comparable performance to other state-of-the-art approaches like vector similarity and the "probabilistic" approach [1]. This paper presents two modest smoothing-based extensions in the LM paradigm.

Whereas the unigram model and other standard approaches to retrieval typically assume bag-of-words independence between terms, modeling even a simple notion of term dependency represents a useful step toward richer modeling of queries and documents. Previous work in bigram modeling provided a valuable first step in this direction within the LM paradigm and demonstrated its empirical merit [16]. Subsequent to this, Dirichlet smoothing with unigram models was found to elegantly and effectively capture the intuition that longer documents should require less smoothing since they provide more support for the maximum-likelihood (ML) estimate [17]. While one would expect bigram models could similarly benefit, we have not seen a Dirichlet-smoothed bigram model described or evaluated in the IR literature. Consequently, we describe such a model here and report on its effectiveness. As with the earlier bigram formulation [16], our approach easily generalizes to higher-order mixtures.

¹ An earlier version of this work was presented in the CLEF 2007 Working Notes.

The second extension we describe addresses smoothing at the collection-level. As suggested above, smoothing plays an important role in inferring accurate document LMs, and it can be accomplished in a principled manner via *maximum a posteriori* (MAP) estimation using a prior model. For IR, the prior is typically estimated from collection statistics, but just as estimating a robust document model is often challenging due to document sparsity, estimating the prior from a small (i.e. sparse) collection can be equally problematic. To address this, we propose estimating the prior from an “expanded” version of the collection containing additional statistics drawn from external corpora. This idea closely parallels previous work expanding documents with similar ones found in external sources [15]. Here, collection-wide statistics are expanded via external corpora to enable more robust estimation of the LM prior. We show simple collection expansion via broad English corpora significantly improves retrieval accuracy.

We evaluated our model and extensions via retrieval experiments on the MALACH archive of automatically transcribed and manually summarized spontaneous speech interviews [9]. These experiments were conducted as part of the Cross-Language Speech Retrieval track’s shared task [11] at the 2007 Cross Language Evaluation Forum. Results show the overall competitive performance of our system as well as the relative contribution of our extensions.

The remainder of our paper is presented as follows: methodology is discussed in §2, relevant details of the MALACH collection and pre-processing are described in §3, evaluation procedure and results are presented in §4, and §5 summarizes and describes future work.

2 Method

2.1 Dirichlet-Smoothed Bigram Modeling

The link recently forged between language modeling and information retrieval [12] established a new mathematical foundation for IR that made a large body of existing theoretical knowledge and empirical experience suddenly applicable. This connection opened the door to an exciting new line of IR research that has already delivered new theoretical insights and excellent empirical results, while at the same time leaving open many interesting directions to pursue.

The core insight of the LM approach is that rather than trying to directly connect a query to its relevant documents by measuring similarity of observed terms, we instead seek an indirect connection by inferring a common underlying stochastic distribution from which query and document arise. The key challenges in this approach are hypothesizing the form of the underlying source models and finding an effective estimation procedure given the brevity of observed evidence.

If we assume *a priori* that all documents are equally likely to be relevant to a given query, then by Bayes inversion we can formulate the document ranking task as estimating query Q ’s likelihood under each document D ’s underlying LM: $P(D|Q) \propto P(Q|D)$. Further assuming complete independence between observed terms (naive Bayes) yields a bag-of-words unigram model in which query

likelihood is estimated by the product of individual term probabilities under the document LM $P(\cdot|D)$.

How do we estimate this model $P(\cdot|D)$? One option is ML. Assuming vocabulary size V , word w_i occurring in D with frequency f_{w_i} , and $P(\cdot|D)$ being parameterized by Θ , we could seek the particular $\hat{\Theta}$ maximizing D 's likelihood

$$P(D|\Theta) = \prod_{i=1}^V \theta_i^{f_{w_i}} \quad (1)$$

which would be the assignment to Θ respecting the empirical frequencies f . However, such use of ML is problematic in that a single unobserved query term would completely nullify query likelihood, making the entire framework exceedingly fragile. The problem here is that in observing only a small sample (i.e. a brief document) from an underlying distribution, effects of chance variation will be prominent and distort sample statistics away from those governing the generating distribution. Fortunately, prior knowledge about the distribution can be leveraged in a principled way via MAP estimation. *A priori*, we might reasonably assume $P(\cdot|D)$ should resemble the collection's *average* document model $P(\cdot|C)$. This, in turn, could be estimated via ML by summing statistics across all documents, which generally do provide sufficient evidence for a robust estimate.

Such prior knowledge can be elegantly incorporated into a language model via the Dirichlet distribution, specified by hyperparameters $\alpha > 0$ and defining a distribution over multinomial parameterizations $P(\Theta; \alpha)$ [6]. For the unigram model defined above, the corresponding Dirichlet prior would be defined as

$$P(\Theta; \alpha) \doteq Dir(\alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^V \theta_i^{\alpha_i - 1} \quad (2)$$

where $Z(\alpha)$ denotes normalization. This prior is particularly convenient for MAP estimation because its distribution is conjugate to the multinomial, meaning the posterior will also be Dirichlet. Hence, combining likelihood (1) and prior (2):

$$P(\Theta|D; \alpha) \propto P(\Theta; \alpha)P(D|\Theta) \propto \prod_{i=1}^V \theta_i^{\alpha_i - 1} \prod_{i=1}^V \theta_i^{f_{w_i}} = \prod_{i=1}^V \theta_i^{f_{w_i} + \alpha_i - 1} \quad (3)$$

A true Bayesian would next compute the predictive distribution over Θ , but we will instead assume a peaked posterior and find the single most-likely $\hat{\Theta}$ to explain our data via the maximum approximation. Comparing our likelihood and posterior equations (1) and (3), we can see that maximizing the posterior is quite similar to maximizing the likelihood, only the data now consists of both the empirical evidence and “pseudo” α observations. In other words, the posterior maximum is simply the combined relative frequency of the observed and pseudo data. Finally, letting $\alpha - 1 = \mu P(\cdot|C)$ for $\mu > 0$, we see our empirical document statistics are smoothed with μ pseudo-counts drawn from our average document model $P(\cdot|C)$ to yield IR's popular Dirichlet-smoothed unigram model [17]

$$P(w|D, C) = \frac{f_w + \mu P(w|C)}{N + \mu} \quad (4)$$

where N specifies the length of D . The attractiveness of this smoothing strategy lies in the fact that as document length increases, providing more evidence for the ML estimate, the impact of the prior model will correspondingly diminish.

To extend this strategy to bigram modeling, we similarly smooth the empirical bigram estimate with hyperparameter μ_1 pseudo-counts distributed fractionally according to the collection prior bigram model, $P(w_i|w_{i-1}, C)$:

$$P(w_i|w_{i-1}, D, C) = \frac{f_{w_{i-1}, w_i} + \mu_1 P(w_i|w_{i-1}, C)}{f_{w_{i-1}} + \mu_1} \quad (5)$$

Unigram and bigram models can then be easily mixed by treating our smoothed unigram distribution $P(w|D, C)$ as an additional prior on the bigram model and adding in μ_2 pseudo-counts drawn from it:

$$P(w_i|w_{i-1}, D, C) = \frac{f_{w_{i-1}, w_i} + \mu_1 P(w_i|w_{i-1}, C) + \mu_2 P(w|D, C)}{f_{w_{i-1}} + \mu_1 + \mu_2} \quad (6)$$

Whereas earlier work inferred the hyperparameters α from data in order to realize a coupled prior tying unigram and bigram models [6], our formulation can be viewed as a less sophisticated alternative that reduces α to three hyperparameters, μ, μ_1 , and μ_2 , to be tuned on development data.

2.2 Collection Expansion

The second extension we describe addresses more robust estimation of the LM prior by performing smoothing at the collection-level. As discussed above, ML estimation of document LMs is hurt by document sparsity, and hence MAP estimation is commonly employed instead using an informative prior induced from the collection. The effectiveness of this strategy, however, relies on accurate estimation of the prior, which can be challenging for small (i.e. sparse) collections.

To address this, we propose estimating the prior from an “expanded” version of the collection containing additional data drawn from external corpora. This approach parallels traditional work in document expansion in which collection documents are expanded with external, related documents [15]. In both cases, the underlying idea of expansion being employed is characteristic of a broad finding in the learning community that having additional similar data enables more robust estimation. In our case of *collection expansion*, we hope to compensate for collection sparsity by drawing upon “similar” data from external corpora.

For this work, we simply leveraged two broad English newspaper corpora: the Wall Street Journal (WSJ) and the North American News Corpus (NANC) [2]. Specifically, we expanded the collection as a linear mixture with 40K sentences (830K words) from WSJ (as found in the Penn Treebank [7]) and 450K sentences (9.5M words) from NANC, with tunable hyperparameters specifying integer mixing ratios between corpora. The particular corpora and mixing scheme used could likely be improved by a more sophisticated strategy. For example, results in §4 show significant improvement for modeling manually-written summaries but not for automatic transcriptions, likely due to mismatch between the external corpora and the automatic transcriptions. Bigram statistics in expansion corpora

were not collected across sentence boundaries, which were manually annotated in WSJ and automatically detected in NANC [8].

3 Data

This section describes the retrieval collection used and pre-processing performed. A more complete description of the collection can be found elsewhere [9,10,11].

Data used came from the Survivors of the Shoah Visual History Foundation (VHF) archive of interviews with Holocaust survivors, rescuers, and witnesses. A subset of this archive was manually and automatically processed by VHF and members of the MALACH initiative (Multilingual Access to Large Spoken Archives) in order to improve access to this archive and other such collections of spontaneous speech content. As part of this effort, interviews were manually segmented and summarized, as well as automatically transcribed (several variant transcriptions were produced). Manual transcription was limited and not provided for interviews included in the retrieval collection. Each interview segment was also manually assigned a set of keywords according to a careful ontology developed by VHF, and two versions of automatically detected keywords were also provided. Topics used for retrieval were based on actual information requests received by VHF from interested parties and were expressed in typical TREC-style with increasingly detailed title, description, and narrative fields [9].

In terms of pre-processing, sentence boundaries were automatically detected to collect more accurate bigram statistics. Boundaries for manual summaries were detected using a standard tool [13] and interview segment keyword phrases were each treated as separate sentences. We noted the presence of multiple contiguous spaces in automatic transcriptions appeared to correlate with sentence-like units (SUs) [3] and so segmented sentences based on them². Use of automatic SU-boundary detection is left for future work [14].

4 Evaluation

This section describes system evaluation, including experimental framework, parameter settings, and results. Retrieval experiments were performed as part of the 2007 Cross Language Evaluation Forum’s Cross-Language Speech Retrieval (CL-SR) task [11].

We used 25 topics for development and 33 for final testing (the 2005 and 2006 CL-SR evaluation sets, respectively; the 2006 test set was re-used for the 2007 evaluation). For the “manual” retrieval condition, segments consisted of manual summaries and keywords. For the “automatic” condition, we used the ASR2006B transcripts and both versions of automatic keywords. Following previous work [17], the unigram Dirichlet smoothing parameter μ was fixed at 2000 for both manual and automatic conditions. Best performance was usually observed with μ_1 set to 1, while optimal μ_2 settings varied.

² Collection documentation does not discuss this.

A limited pseudo-relevance feedback (PRF) scheme was also employed. As in standard practice, documents were ranked by the model according to the original query, with the most likely documents taken to comprise its feedback set (the number of feedback documents used varied). The query was then reformulated by adding the 50 most frequent bigrams from each feedback document. A tuning parameter specified a multiplier for the original query counts to provide a means of weighting the original query relative to the feedback set. This scheme likely could be improved by separate treatment for unigram feedback and weighting feedback documents by document likelihood under the original query.

Results in Table 1 show performance of our five official runs on development and test sets³; queries used were: title-only (T), title and description (TD), and title, description, and narrative (TDN). Representative strong results achieved in 2007’s and previous years’ CL-SR tracks [10,11] are also shown, though it should be noted that our results on the development set correspond to tuning on those queries whereas the CL-SR’05 official results do not. Retrieval accuracy was measured using mean-average precision reported by `trec_eval` version 8.1⁴.

Table 1. Mean-average precision retrieval accuracy of submitted runs. CL-SR columns indicate representative strong results achieved in that year’s track on the same query set [10,11]. Runs marked above with +/- were reported in the 2007 track report to represent statistical significance and non-significance, respectively.

Collection	Queries	Dev	CL-SR’05	Test	CL-SR’06	CL-SR’07
Manual	TDN	.3829	-	.2870	.2902	.2847
	TD	.3443	.3129	.2366+	.2710	.2761+
	T	.3161	-	.2348	.2489	-
Auto	TDN	.1623	.2176	.0910	.0768	-
	TD	.1397	.1653	.0785-	.0754	.0855-

Table 2 shows the impact of our extensions compared to the baseline Dirichlet-smoothed unigram retrieval model for the no-PRF “manual” condition. Of the two extensions, collection expansion is seen to have greater effect, with the combination yielding the best result. The effect of the extensions with the “automatic” condition was marginal (the best absolute improvement seen was 0.3% achieved by the bigram model). With collection expansion, we suspect this is due to the mismatch between the collection’s spontaneous speech and the text corpora used for expansion (§2), and we plan to investigate use of better matched corpora in future work. As for the bigram model, automatic transcription noise is more problematic than with unigrams since recognition error further impacts prediction of subsequent terms. One strategy for addressing this would be to work off the recognition lattice instead of the one-best transcription. Another challenge to the bigram model is the presence of disfluency in spontaneous speech, which

³ Following submission of official runs, we found a bug affecting our parsing of the *narrative* field of three test queries. Table 1 show system performance with the bug fixed. Without the fix, **Manual**-TDN on the test set was .2577 and **Auto**-TDN was .0831.

⁴ http://trec.nist.gov/trec_eval

Table 2. Improvement in mean-average precision on the development set over the unigram baseline model for Dirichlet-smoothed bigram modeling and collection expansions, alone and in combination (manual condition, no pseudo-relevance feedback)

Model	T	TD	TDN
Unigram baseline	.2605	.2722	.2810
Dirichlet bigram	.2545 (-2.3%)	.2852 (4.8%)	.2967 (5.6%)
Collection Expansion	.2716 (4.3%)	.3021 (11.0%)	.3236 (15.2%)
Combination	.2721 (4.5%)	.3091 (13.6%)	.3369 (19.9%)

disrupts bigram statistics. Automatic detection and deletion of disfluency could help address this and thereby also render the spoken document more amenable to smoothing via external text corpora [5].

For manual retrieval with PRF, the combination of extensions was used in selecting the set of documents for feedback. For PRF runs using this feedback set, the extensions were seen to provide minimal further benefit, with PRF tuning parameters dominating the variance in performance observed. Since PRF produces a query more tailored to collection statistics, expanded collection statistics may be less useful in PRF settings.

5 Conclusion and Future Work

This paper presented two smoothing-based extensions to the standard language model approach to information retrieval: Dirichlet-smoothed bigram modeling and collection expansion. Empirical results demonstrated the relative contribution of the extensions and competitive overall system performance.

Future work will explore two lines of research in LM-based information retrieval [4]: inferring latent structure to derive richer representations for modeling, and revisiting existing SDR retrieval methodology with greater attention to modeling spontaneous speech phenomena.

Acknowledgments

The authors thank Will Headden and our anonymous reviewers for their valuable comments. This work was initiated by the first author while hosted at the Institute of Formal and Applied Linguistics (ÚFAL) at Charles University in Prague. Support for this work was provided by NSF PIRE Grant No OISE-0530118 and DARPA GALE contract HR0011-06-2-0001. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the supporting agencies.

References

1. Fang, H., Tao, T., Zhai, C.X.: A formal study of information retrieval heuristics. In: SIGIR 2004: Proc. 27th ACM SIGIR conference, pp. 49–56 (2004)
2. David Graff. North American News Text Corpus, Linguistic Data Consortium. LDC95T21 (1995)

3. LDC. Simple metadata annotation specification 6.2. Technical report (2004)
4. Lease, M.: Natural language processing for information retrieval: the time is ripe (again). In: *Proceedings of the 1st Ph.D. Workshop at the ACM Conference on Information and Knowledge Management (PIKM)* (2007)
5. Lease, M., Johnson, M., Charniak, E.: Recognizing disfluencies in conversational speech. *IEEE Transactions on Audio, Speech and Language Processing* 14(5), 1566–1573 (2006)
6. MacKay, D.J.C., Peto, L.: A hierarchical Dirichlet language model. *Natural Language Engineering* 1(3), 1–19 (1995)
7. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330 (1993)
8. McClosky, D., Charniak, E., Johnson, M.: Effective self-training for parsing. In: *Proceedings of the HLT-NAACL conference*, pp. 152–159 (2006)
9. 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: *SIGIR 2004: Proc. of the 27th annual international ACM SIGIR conference*, pp. 41–48 (2004)
10. Oard, D.W., 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.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 744–758. Springer, Heidelberg* (2007)
11. Pecina, P., Hoffmannova, P., Jones, G.J.F., Zhang, Y., Oard, D.W.: Overview of the CLEF-2007 cross language speech retrieval track. In: *Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 674–686. Springer, Heidelberg* (2008)
12. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: *Proceedings of the 21st annual international ACM SIGIR conference*, pp. 275–281 (1998)
13. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: *Proceedings of the fifth conference on Applied natural language processing*, pp. 16–19 (1997)
14. Roark, B., Liu, Y., Harper, M., Stewart, R., Lease, M., Snover, M., Shafran, I., Dorr, B., Hale, J., Krasnyanskaya, A., Yung, L.: Reranking for sentence boundary detection in conversational speech. In: *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 545–548 (2006)
15. Singhal, A., Pereira, F.: Document expansion for speech retrieval. In: *Proc. of the 22nd annual international ACM SIGIR conference*, pp. 34–41 (1999)
16. Song, F., Croft, W.B.: A general language model for information retrieval. In: *Proceedings of the eighth international conference on Information and knowledge management (CIKM)*, pp. 316–321 (1999)
17. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. *ACM. Trans. Inf. Syst.* 22(2), 179–214 (2004)

Model Fusion Experiments for the CLSR Task at CLEF 2007

Muath Alzghool and Diana Inkpen

School of Information Technology and Engineering
University of Ottawa
{alzghool,diana}@site.uottawa.ca

Abstract. This paper presents the participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) task at CLEF 2007. We present the results of the submitted runs for the English collection. We have used two Information Retrieval systems in our experiments: SMART and Terrier, with two query expansion techniques: one based on a thesaurus and the second one based on blind relevant feedback. We proposed two novel data fusion methods for merging the results of several models (retrieval schemes available in SMART and Terrier). Our experiments showed that the combination of query expansion methods and data fusion methods helps to improve the retrieval performance. We also present cross-language experiments, where the queries are automatically translated by combining the results of several online machine translation tools. Experiments on indexing the manual summaries and keywords gave the best retrieval results.

Keywords: Data Fusion, Retrieval Models, Query Expansion.

1 Introduction

This paper presents the third participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) track, at CLEF 2007. We present our systems, followed by results for the submitted runs for the English 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. Several researchers in the literature have explored the idea of combining the results of different retrieval strategies, different document representations and different query representations; the motivation is that each technique will retrieve different sets of relevant documents; therefore combining the results could produce a better result than any of the individual techniques. We propose new data fusion techniques for combining the results of different Information Retrieval (IR) schemes. We applied our data fusion techniques to monolingual settings and to cross-language settings where the queries are automatically translated from French and Spanish into English by combining the results of several online machine translation (MT) tools. At the end we present the best results, when manual summaries and manual keywords were indexed.

2 System Description

The University of Ottawa Cross-Language Information Retrieval systems were built with off-the-shelf components. For the retrieval part, the SMART [3, 11] IR system and the Terrier [2, 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 *nnn.ntn*, *ntn.ntn*, *lnn.ntn*, *ann.ntn*, *ltn.ntn*, *atn.ntn*, *ntn.nnn*, *nnc.ntc*, *ntc.ntc*, *ntc.nnc*, *lnc.ntc*, *anc.ntc*, *ltc.ntc*, *atc.ntc* weighting schemes [3, 11]; *lnn.ntn* performs very well in CLEF-CLSR 2005 and 2006 [1, 6].

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 [2, 9]. We experimented with the *In(exp)C2* weighting model, one of Terrier's DFR-based document weighting models.

For translating the queries from French and Spanish 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 [6] were used for translating from Spanish and from French into English. 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.

We have used two query expansion methods. The first one is based on the Shoah Visual History Foundation thesaurus provided with the Mallach collection; our method adds two items and their alternatives (synonyms) from the thesaurus, based on the similarity between the thesaurus terms and the title field for each topic. More specifically, to select two items from the thesaurus, we used SMART with the title of each topic as query and the thesaurus terms as documents, using the weighting scheme *lnn.ntn*. After computing the similarity, the top two thesaurus terms were added to the topic; for these terms all the alternative terms was also added to the topic. For example, in topic 3005, the title is "Death marches", and the most similar terms from the thesaurus are "death marches" and "deaths during forced marches"; the alternative terms for theses terms are "death march" and "Todesmärsche". Table 1 shows two entries from the thesaurus; each entry contains six types of fields: name – contains a unique numeric code for each entry, label – a phrase or word which represents the entry, alt-label – contains the alternative phrase or the synonym for the entry, usage – contains the usage or the definition of the entry. There are two more relations in the thesaurus: *is-a* and *of-type*, which contain the numeric code of the entry involved in the relation. The second query expansion method extracts the most informative terms from the top-returned documents as the expanded query terms. In this expansion process, 12 terms from the returned documents (the top 15 documents) were added to the topic, based on Bose-Einstein 1 model (Bo1) [4, 9]; we have put a restriction on the new terms: their document frequency must be less than the

Table 1. The top two entries from the thesaurus that are similar to the topic title “Death marches”

```

<keyword>
  <name>9125</name>
  <alt-label>death march</alt-label>
  <alt-label>Todesmärsche</alt-label>
  <broadener-term>15445</broadener-term>
  <label>death marches</label>
  <of-type>5289</of-type>
  <usage>Forced marches of prisoners over long distances,
under heavy guard and extremely harsh conditions. (The term
was probably coined by concentration camp prisoners.)</usage>
</keyword>
<keyword>
  <name>15460</name>
  <broadener-term>15445</broadener-term>
  <label>deaths during forced marches</label>
  <of-type>4109</of-type>
  <usage>The daily experience of individuals with death
during forced marches that was not the result of executions,
punishments, arbitrary killings or suicides.</usage>
</keyword>

```

maximum document frequency in the title of the topic. The aim of this restriction is avoid more-general terms being added to the topic. Any term that satisfies this restriction will be a part of the new topic. We have also up weighted the title terms five times higher than the other terms in the topic.

For the data fusion part, we proposed two methods that use the sum of normalized weighted similarity scores of 15 different IR schemes as shown in the following formulas :

$$Fusion1 = \sum_{i \in IR \text{ schemes}} [W_r^4(i) + W_{MAP}^3(i)] * NormSim_i \quad (1)$$

$$Fusion2 = \sum_{i \in IR \text{ schemes}} W_r^4(i) * W_{MAP}^3(i) * NormSim_i \quad (2)$$

where $W_r(i)$ and $W_{MAP}(i)$ are experimentally determined weights based on the recall (the number of relevant documents retrieved) and precision (MAP score) values for each IR scheme computed on the training data. For example, suppose that two retrieval runs $r1$ and $r2$ give 0.3 and 0.2 (respectively) as MAP scores on training data; we normalize these scores by dividing them by the maximum MAP value: then $W_{MAP}(r1)$ is 1 and $W_{MAP}(r2)$ is 0.66 (then we compute the power 3 of these weights, so that one weight stays 1 and the other one decreases; we chose power 3 for MAP score and power 4 for recall, because the MAP is more important than the recall). We

hope that when we multiply the similarity values with the weights and take the summation over all the runs, the performance of the combined run will improve. $NormSim_i$ is the normalized similarity for each IR scheme. We did the normalization by dividing the similarity by the maximum similarity in the run. The normalization is necessary because different weighting schemes will generate different range of similarity values, so a normalization method should be applied to each run. Our method differed from the work done by Fox and Shaw in 1994 [5] and Lee in 1995 [7]; they combined the results by taking the summation of the similarity scores without giving any weight to each run. In our work we weight each run according to the precision and recall on the training data.

3 Experimental Results

3.1 Submitted Runs

Table 2 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) and Recall. 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; [8,9] give more information about the training and test data. For the `uoEnTDtManF1` and `uoEnTDtQExF1` runs we used the Fusion1 formula for data fusion; and for `uoEnTDtQExF2`, `uoFrTDtF2`, and `uoEsTDtF2` we used the Fusion2 formula for data fusion. We used blind relevance feedback and query expansion from the thesaurus for the `uoEnTDtManF1`, `uoEnTDtQExF1`, and `uoEnTDtQExF2` runs; we didn't use any query expansion techniques for `uoFrTDtF2` and `uoEsTDtF2`.

Our required run, English TD, obtained a MAP score of 0.0855. Comparing this result to the median and average of all runs submitted by all the teams that participated in the

Table 2. Results of the five submitted runs, for topics in English, French, and Spanish. The required run (English, title + description) is in bold.

Runs	MAP	Recall	Description
uoEnTDtManF1	0.2761	1832	English: Fusion 1, query expansion methods, fields: MANUALKEYWORD + SUMMARY
uoEnTDtQExF1	0.0855	1333	English: Fusion 1, query expansion methods, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoEnTDtQExF2	0.0841	1336	English: Fusion 2, query expansion methods, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoFrTDtF2	0.0603	1098	French : Fusion 2, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2
uoEsTDtF2	0.0619	1171	Spanish : Fusion 2, fields: ASRTEXT2004A + AUTOKEYWORD2004A1, A2

track (0.0673, 0.0785) [10], our result was significantly better (based on atwo-tailed Wilcoxon Signed-Rank Test for paired samples at $p < 0.05$ across the 33evaluation topics) with a relative improvement of 21% and 8 %; there is a small improvement using Fusion1 (uoEnTDtQExF1) over Fusion2 (uoEnTDtQExF2), but this improvement is not significant.-

3.2 Comparison of Systems and Query Expansion Methods

In order to compare between different methods of query expansion and a base run without query expansion, we selected the base run with the weighting scheme lnn.ntn, topic fields title and description, and document fields ASRTEXT2004A, AUTOKEYWORD2004A1, and AUTOKEYWORD2004A2. We used the two techniques for query expansion, one based on the thesaurus and the other one on blind relevance feedback (denoted Bo1 in Table 3). We present the results (MAP scores) with and without query expansion, and with the combination of both query expansion methods, on the test and training topics. According to Table 3, we note that both methods help to improve the retrieval results, but the improvement is not significant on the training and test data; also the combination of the two methods helps to improve the MAP score on the training data (not significantly), but not on the test data.

Table 3. Results (MAP scores) for Terrier and SMART, with or without relevance feedback, for English topics (using the TD query fields)

	System	Training	Test
1	lnn.ntn	0.0906	0.0725
2	lnn.ntn +thesaurus	0.0941	0.0730
3	lnn.ntn +Bo1	0.0954	0.0811
4	lnn.ntn+ thesaurus+ Bo1	0.0969	0.0799

3.3 Experiments Using Data Fusion

We applied the data fusion methods described in section 2 to 14 runs produced by SMART and one run produced by Terrier; all runs was produced using a combination of the two methods of query expansion as described in section 2. Performance results for each single run and fused runs are presented in Table 4, in which % change is given with respect to the run providing better effectiveness in each combination on the training data. The Manual English column represents the results when only the manual keywords and the manual summaries were used for indexing the documents using English topics, the Auto-English column represents the results when automatic fields are indexed from the documents (ASRTEXT2004A, and AUTOKEYWORD2004A1, A2) using English topics. For cross-languages experiments the results are represented in the columns Auto-French, and Auto-Spanish.

Data fusion helps to improve the performance (MAP score) on the test data The best improvement using data fusion (Fusion1) was on the French cross-language

experiments with 21.7%, which is statistically significant while on monolingual the improvement was only 6.5% which is not significant. Also, there is an improvement in the number of relevant documents retrieved (recall) for all the experiments, except Auto-French on the test data, as shown in Table 5. We computed these improvements relative to the results of the best single-model run, as measured on the training data. This supports our claim that data fusion improves the recall by bringing some new documents that were not retrieved by all the runs. On the training data, the Fusion2 method gives better results than Fusion1 for all cases except on Manual English, but on the test data Fusion1 is better than Fusion2. In general, the data fusion seems to help, because the performance on the test data is not always good for weighting schemes that obtain good results on the training data, but combining models allows the best-performing weighting schemes to be taken into consideration.

The retrieval results for the translations from French were very close to the monolingual English results, especially on the training data, but on the test data the difference was significantly worse. For Spanish, the difference was significantly worse on the training data, but not on the test data.

Experiments on manual keywords and manual summaries showed high improvements, the MAP score jumped from 0.0855 to 0.2761 on the test data.

Table 4. Results (MAP scores) for 15 weighting schemes using Smart and Terrier (the In(exp)C2 model), and the results for the two Fusions Methods. In bold are the best scores for the 15 single runs on the training data and the corresponding results on the test data. Underlined are the results of the submitted runs.

Weighting scheme	Manual English		Auto-English		Auto-French		Auto-Spanish	
	Train.	Test	Train.	Test	Train.	Test	Train.	Test
nnc.ntc	0.2546	0.2293	0.0888	0.0819	0.0792	0.055	0.0593	0.0614
ntc.ntc	0.2592	0.2332	0.0892	0.0794	0.0841	0.0519	0.0663	0.0545
lnc.ntc	0.2710	0.2363	0.0898	0.0791	0.0858	0.0576	0.0652	0.0604
ntc.nnc	0.2344	0.2172	0.0858	0.0769	0.0745	0.0466	0.0585	0.062
anc.ntc	0.2759	0.2343	0.0723	0.0623	0.0664	0.0376	0.0518	0.0398
ltc.ntc	0.2639	0.2273	0.0794	0.0623	0.0754	0.0449	0.0596	0.0428
atc.ntc	0.2606	0.2184	0.0592	0.0477	0.0525	0.0287	0.0437	0.0304
nnn.ntn	0.2476	0.2228	0.0900	0.0852	0.0799	0.0503	0.0599	0.061
ntn.ntn	0.2738	0.2369	0.0933	0.0795	0.0843	0.0507	0.0691	0.0578
lnn.ntn	0.2858	0.245	0.0969	0.0799	0.0905	0.0566	0.0701	0.0589
ntn.nnn	0.2476	0.2228	0.0900	0.0852	0.0799	0.0503	0.0599	0.061
ann.ntn	0.2903	0.2441	0.0750	0.0670	0.0743	0.038	0.057	0.0383
ltn.ntn	0.2870	0.2435	0.0799	0.0655	0.0871	0.0522	0.0701	0.0501
atn.ntn	0.2843	0.2364	0.0620	0.0546	0.0722	0.0347	0.0586	0.0355
In(exp)C2	0.3177	0.2737	0.0885	0.0744	0.0908	0.0487	0.0747	0.0614
Fusion 1	0.3208	<u>0.2761</u>	0.0969	<u>0.0855</u>	0.0912	0.0622	0.0731	0.0682
% change	1.0%	<u>0.9%</u>	0.0%	<u>6.5%</u>	0.4%	21.7%	-2.2%	10.0%
Fusion 2	0.3182	0.2741	0.0975	0.0842	0.0942	<u>0.0602</u>	0.0752	0.0619
% change	0.2%	0.1%	0.6%	5.1%	3.6%	<u>19.1%</u>	0.7%	0.8%

Table 5. Results (number of relevant documents retrieved) for 15 weighting schemes using Terrier and SMART, and the results for the Fusions Methods. In bold are the best scores for the 15 single runs on training data and the corresponding test data; underlined are the submitted run.

Weighting scheme	Manual English		Auto-English		Auto- French		Auto- Spanish	
	Train.	Test	Train.	Test	Train.	Test	Train.	Test
nnc.ntc	2371	1827	1726	1306	1687	1122	1562	1178
ntc.ntc	2402	1857	1675	1278	1589	1074	1466	1155
lnc.ntc	2402	1840	1649	1301	1628	1111	1532	1196
ntc.nnc	2354	1810	1709	1287	1662	1121	1564	1182
anc.ntc	2405	1858	1567	1192	1482	1036	1360	1074
ltc.ntc	2401	1864	1571	1211	1455	1046	1384	1097
atc.ntc	2387	1858	1435	1081	1361	945	1255	1011
nnn.ntn	2370	1823	1740	1321	1748	1158	1643	1190
ntn.ntn	2432	1863	1709	1314	1627	1093	1502	1174
lnn.ntn	2414	1846	1681	1325	1652	1130	1546	1194
ntn.nnn	2370	1823	1740	1321	1748	1158	1643	1190
ann.ntn	2427	1859	1577	1198	1473	1027	1365	1060
ltn.ntn	2433	1876	1582	1215	1478	1070	1408	1134
atn.ntn	2442	1859	1455	1101	1390	975	1297	1037
In(exp)C2	2638	1823	1624	1286	1676	1061	1631	1172
Fusion 1	2645	1832	1745	1334	1759	1147	1645	1219
% change	0.3%	0.5 %	0.3%	1.0%	0.6%	-1.0%	0.1%	2.4%
Fusion 2	2647	1823	1727	1337	1736	1098	1631	1172
% change	0.3%	0.0%	0.8%	1.2%	-0.7%	-5.5%	-0.7%	-1.5%

4 Conclusion

We experimented with two different systems: Terrier and SMART, with combining the various weighting schemes for indexing the document and query terms. We proposed two approaches for query expansion, one based on the thesaurus and another one based on blind relevance feedback. The combination of the query expansion methods obtained a small improvement on the training and test data (not statistically significant according to a Wilcoxon signed test).

Our focus this year was on data fusion: we proposed two methods to combine different weighting scheme from different systems, based on weighted summation of normalized similarity measures; the weight for each scheme was based on the relative precision and recall on the training data. Data fusion helps to improve the retrieval significantly for some experiments (Auto-French) and for other not significantly (Manual English).

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.

Combining query expansion methods and data fusion helped to improve the retrieval significantly comparing to the median and average of all required runs submitted by all the teams that participated in the track.

In future work we plan to investigate more methods of data fusion, removing or correcting some of the speech recognition errors in the ASR content words, and to use speech lattices for indexing.

References

1. Alzghool, M., Inkpen, D.: Experiments for the Cross Language Speech Retrieval Task at CLEF 2006. In: *Proceedings of CLEF 2006*. LNCS, vol. 4730, pp. 778–785. Springer, Heidelberg (2006)
2. 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)
3. Buckley, C., Salton, G., Allan, J.: Automatic retrieval with locality information using SMART. In: *Text REtrieval Conference (TREC-1)*, March, pp. 59–72 (1993)
4. 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)
5. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: *Proceedings of the Third Text REtrieval Conference (TREC-3)*, vol. 500–215, National Institute of Standards and Technology Special Publication (1994)
6. 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)
7. Lee, J.H.: Combining multiple evidence from different properties of weighting schemes. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 180–188 (1995)
8. 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)*
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. Pecina, P., Hoffmannová, P., Jones, G.J.F., Zhang, Y., Oard, D.W.: Overview of the CLEF-2007 Cross Language Speech Retrieval Track. In: *Working Notes of the CLEF 2007 Evaluation*, Budapest, Hungary, 11 pages (2007)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic retrieval. *Information Processing and Management* 24(5), 513–523 (1988)

Dublin City University at CLEF 2007: Cross-Language Speech Retrieval Experiments

Ying Zhang, Gareth J.F. Jones, and Ke Zhang

Centre for Digital Video Processing & School of Computing
Dublin City University, Dublin 9, Ireland
{yzhang,gjones,kzhang}@computing.dcu.ie

Abstract. The Dublin City University participation in the CLEF 2007 CL-SR English task concentrated primarily on issues of topic translation. Our retrieval system used the BM25F model and pseudo relevance feedback. Topics were translated into English using the Yahoo! BabelFish free online service combined with domain-specific translation lexicons gathered automatically from *Wikipedia*. We explored alternative topic translation methods using these resources. Our results indicate that extending machine translation tools using automatically generated domain-specific translation lexicons can provide improved CLIR effectiveness for this task.

1 Introduction

The Dublin City University participation in the CLEF 2007 CL-SR task focused on extending our CLEF 2006 system [1] to investigate combinations of general and domain-specific topic translation resources. Our 2006 system used the BM25F field combination approach [2] with summary-based pseudo relevance feedback (PRF) [3]. Our submissions to CLEF 2007 included both English monolingual and French–English bilingual tasks using automatic only and combined automatic and manual fields. The Yahoo! BabelFish machine translation system¹ was used for baseline topic translation into English; this was then combined with domain-specific translation lexicons gathered automatically from *Wikipedia*.

The remainder of this paper is structured as follows: Section 2 summarises the features of the BM25F retrieval model, Section 3 overviews our retrieval system, Section 4 describes our topic translation methods, Section 5 presents our experimental results, and Section 6 concludes the paper with a discussion of our results.

2 Field Combination

The English collection comprises 8104 “documents” taken from 589 hours of speech data. The spoken documents are provided with a rich set of data fields,

¹ babelfish.yahoo.com

full details of these are given in [4,5]. In this work, we explored field combination based on the following fields: an automatic transcription of the spoken content (the ASR2006B field); automatically generated keywords (AKW1,AKW2); manually generated keywords (MK); manual summary of each segment (SUM); and names of all individuals appearing in the segment.

[2] demonstrates the weaknesses of standard combination methods and proposes 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 weighted summation of the multiple fields of the documents to form a single field for each document. The importance of each document field for retrieval is determined empirically, each field is then multiplied by a scalar constant representing the importance of this field, and the components of all fields are then summed.

3 Okapi Retrieval System

The basis of our experimental system is the City University research distribution version of the Okapi system [6]. The documents and search topics are processed to remove stopwords from a standard list of about 260 words, suffix stripped using the Okapi implementation of Porter stemming [7] and terms are indexed using a small standard set of synonyms. None of these procedures were adapted for the CLEF 2007 CL-SR test collection.

Document terms were weighted using the Okapi BM25 weighting scheme,

$$cw(i, j) = cfw(i) \times \frac{tf(i, j) \times (k_1 + 1)}{k_1 \times ((1 - b) + (b \times ndl(j))) + tf(i, j)}$$

$$cfw(i) = \log \left(\frac{(rload + 0.5)(N - n(i) - bigrload + rload + 0.5)}{(n(i) - rload + 0.5)(bigrload - rload + 0.5)} \right)$$

where $cw(i, j)$ = the weight of term i in document j ; $n(i)$ = total number of documents containing term i ; N = total number of documents in the collection; $tf(i, j)$ = within document term frequency; $dl(j)$ = length of j ; $avgdl$ = average document length in the collection; $ndl(j) = dl(j)/avgdl$ is the normalized document length; and k_1 and b are empirically-tuned constants for a particular collection. $bigrload$ is an assumed number of relevant documents and $rload$ the number of these containing i . These take the standard values of 4 and 5 respectively [8]. The matching score for each document is computed by summing the weights of terms appearing in the query and the document. In this investigation we use the summary-based PRF method for query-expansion described in [3].

4 MT-Based Query Translation

Machine Translation (MT) based query translation using an existing MT system has been widely used in cross-language information retrieval (CLIR) with good average performance. In our experiments, topics were translated into English

using the Yahoo! BabelFish system powered by SYSTRAN. While BabelFish can provide reasonable translations for general language expressions, it is not sufficient for domain-specific terms such as personal names, organization names, place names, etc. To reduce the errors introduced by such terms during query translation, we augmented the standard BabelFish with domain-specific lexicon resources gathered from *Wikipedia*².

4.1 Domain-Specific Lexicon Construction

As a multilingual hypertext medium, Wikipedia has been shown to be a valuable new source of translation information [9,10]. Unlike the web, the hyperlinks in Wikipedia have a more consistent pattern and meaningful interpretation. A Wikipedia page written in one language can contain hyperlinks to its counterparts in other languages, where the hyperlink basenames are translation pairs. For example, the English wikipedia page en.wikipedia.org/wiki/World_War_II contains hyperlinks to German de.wikipedia.org/wiki/Zweiter_Weltkrieg, French fr.wikipedia.org/wiki/Seconde_Guerre_mondiale, and Spanish es.wikipedia.org/wiki/Segunda_Guerra_Mundial. The English term “World War II” is the translation of the German term “Zweiter Weltkrieg”, the French term “Seconde Guerre mondiale”, and the Spanish term “Segunda Guerra Mundial”.

Additionally, we observed that multiple English wikipedia URLs en.wikipedia.org/wiki/World_War_II, en.wikipedia.org/wiki/World_War_2, en.wikipedia.org/wiki/WW2, and en.wikipedia.org/wiki/Second_world_war are redirected to the same wikipedia page and the URL basenames “World War II”, “World War 2”, “WW2”, and “Second world war” are synonyms. Using all these English terms during query translation is a straightforward approach to automatic post-translation query expansion.

To utilize the multilingual linkage and the link redirection features, we implemented a three-stage automatic process to extract German, French, and Spanish to English translations from Wikipedia:

1. An English vocabulary list was constructed by performing a limited crawl of the English wikipedia³, Category:World War II. This category is likely to contain links to pages and subcategories relevant to entities appearing in the document collection. In total, we collected 7431 English web pages.
2. For each English page, we extracted the hyperlinks to each of the query languages. This provided a total of 4446, 3338, and 4062 hyperlinks to German, Spanish, and French, respectively.
3. We then selected the basenames of each pair of hyperlinks (German–English, French–English, and Spanish–English) as translations and added them into our domain-specific lexicons. Non-English multi-word terms were added into the phrase dictionary for each query language. These phrase dictionaries are later used for phrase identification during query pre-processing.

² www.wikipedia.org

³ <http://en.wikipedia.org>



Fig. 1. An example of French–English query translation. (Topic numbered 3005).

4.2 Query Translation Process

As shown in Figure 1, our query translation process is performed as follows:

1. Query pre-processing: We used the phrase dictionary with the maximum forward matching algorithm to segment each query Q into a list of terms $\{q_1, q_2, q_3, \dots, q_n\}$.
2. Domain-specific lexicon lookup: For each query term q_i (where $i \in (1, n)$), we obtained all its English translations $\{e_{i1}, e_{i2}, e_{i3}, \dots, e_{im}\}$ via a domain-specific lexicon look-up.
3. BabelFish translation: we then translated the original query Q into the English query E using the Yahoo! BabelFish system.
4. Translation results merging: For each English term e_{ij} (where $i \in (1, n)$ and $j \in (1, m)$) obtained in Step 2, we appended it to the end of the translated English query E .

5 Experimental Results

In this section we report results for combinations of manual only fields, automatic only fields and combining both manual and automatic fields. Monolingual retrieval results show precision at cutoff ranks of 5, 10 and 30, standard TREC mean average precision (MAP) and recall in terms of the total number of relevant documents retrieved for the test topic set. CLIR results compare alternative topic translations resources showing MAP and precision at rank 10. Runs formally submitted for evaluation are indicated by an asterisk in the tables.

Table 1. Results for English monolingual retrieval. No significant difference observed.

RUN	Description	Query Fields	Recall	MAP	P@5	P@10	P@30
<i>Manual field combination</i>							
(MK $\times$ 1+SUM $\times$ 1, $k_1 = 1.0$, $b = 0.5$)							
Baseline	TDN	1850	0.2773	0.4909	0.4576	0.4182	
*PRF	TDN	1903	0.2847	0.4970	0.4515	0.4222	
<i>Automatic field combination</i>							
(AK1 $\times$ 1+AK2 $\times$ 1+ASR2006B $\times$ 2, $k_1 = 8.0$, $b = 0.5$)							
Baseline	TD	1311	0.0735	0.1697	0.1697	0.1677	
*PRF	TD	1360	0.0787	0.1697	0.1727	0.1636	
<i>Manual and automatic field combination</i>							
(MK $\times$ 4+SUM $\times$ 4+ASR2006B $\times$ 1, $k_1 = 3.0$, $b = 0.6$)							
*Baseline	TD	1907	0.2399	0.4364	0.3818	0.3838	
*PRF	TD	1974	0.2459	0.4364	0.3818	0.3556	

5.1 System Parameters

Our retrieval system requires a number of parameters to be set for the term weighting, field combination, and PRF components. All parameter values were set empirically using the 63 English language CLEF 2007 CL-SR training topics.

Term weighting and field combination. Based on the training runs the term weighting and field combination parameters were set as follows: *Manual data field combination*, Okapi parameters $k_1 = 1.0$ and $b = 0.5$ with document fields weighted as MK $\times$ 1, and SUM $\times$ 1; *Automatic data field combination*, $k_1 = 8.0$ and $b = 0.5$ document fields weighted as A1K $\times$ 1, AK2 $\times$ 1, and ASR06B $\times$ 2; and *Manual and automatic data field combination*, $k_1 = 3.0$ and $b = 0.6$ document fields weighted as MK $\times$ 4, SUM $\times$ 4, and ASR06B $\times$ 1.

Pseudo-relevance feedback. The top t ranked expansion terms taken from document summaries are added to the original query. The original topic terms are up-weighted by a factor α relative to the expansion terms. Our PRF query expansion involves five parameters as follows: t = number of the expansion terms selected; s = number of sentences selected as the document summary; d_1 = number of documents used for sentence selection; d_2 = is the number of documents used for expansion term ranking; α = the up-weighting factor.

PRF parameter selection is not necessarily consistent from one collection (indexed using different field combination methods) to another. Our experiments showed that $t = 60$, $s = 6$, $d_1 = 3$, $d_2 = 20$, and $\alpha = 3.0$ give the best results for the manual data field combination, and manual and automatic data field combination; while $t = 40$, $s = 6$, $d_1 = 3$, $d_2 = 20$, and $\alpha = 3.0$ produce the best results for the automatic data field combination.

Table 2. Results for cross-lingual retrieval. (TD runs on automatic field combination, $A1K \times 1 + AK2 \times 1 + ASR2006B \times 2$, $k_1 = 8.0$, $b = 0.5$. * shows significance at 0.05 level.)

RUN Description	*French		Spanish		German	
	MAP	P@10	MAP	P@10	MAP	P@10
BabelFish baseline	0.0476	0.1242	0.0566	0.1364	0.0563	0.1394
BabelFish+PRF	0.0501	0.1242	0.0541	0.1303	0.0655	0.1303
BabelFish+LEX	0.0606*	0.1394	0.0581	0.1394	0.0586	0.1424
BabelFish+LEX+PRF	0.0636	0.1394	0.0588	0.1273	0.0617	0.1364

5.2 Field Combination and Summary-Based PRF

Table 1 shows monolingual English retrieval results for both the baseline condition without application of PRF and with summary-based PRF. For the combination of the MK and SUM fields it can be seen that application of PRF generally produces a small improvement in performance. Note that the topics here use all three topics fields Title, Description and Narrative (TDN), and thus these results cannot be compared directly to any other results shown here which use only Title and Description fields (TD). Similarly for both the automatic only fields runs combining AK1, AK2 and ASR2006B, and the combination of manual and automatic fields using MK, SUM and ASR2006B, application of PRF produces a small improvement in average and high rank precision, although there appear to be some problems at lower ranks which we intend to investigate.

5.3 Yahoo! BabelFish Combined with Domain-Specific Lexicons

Table 2 shows CLIR results for standard BabelFish translation (BabelFish baseline), and augmented translations using the domain-specific lexicons (BabelFish+LEX). The BabelFish+LEX method led to a significant improvement (27%) for the French–English retrieval task, but only 3% and 4% in Spanish–English and German–English, respectively. This can be explained by the fact that the MAP values for the baseline runs of German and Spanish are much higher than the MAP for the French baseline. We noticed that the description field of German topics sometimes contains additional explanation enclosed by square brackets. The effect of this was often that more correct documents are retrieved in the German–English task. We therefore believe that the BabelFish system gives a better translation from Spanish, rather than French and German, to English.

At the individual query level (shown in Table 3), we observe that retrieval effectiveness sometimes degrades slightly when the query is augmented to include translations from our domain-specific lexicons, despite the fact that they are correct translations of the original query terms. This occurred mainly due to the fact that additional terms result in a decrease in the rank of relevant documents because they are too general within the collection. For example, “war”, “Europe”, “Poland”, “holocaust”, “country”, “people”, “history”, etc.

We used the summary-based PRF to provide post-translation query expansion in all CLIR runs (see BabelFish+PRF and BabelFish+LEX +PRF shown in

Table 3. Examples of using extra translations from the domain-specific lexicons led to a deterioration in retrieval effectiveness. (TD runs on automatic field combination, $A1K \times 1 + AK2 \times 1 + ASR06B \times 2$, $k_1 = 8.0$, $b = 0.5$.)

Query ID	MAP		Additional Translations from Lexicons
	BabelFish	BabelFish+LEX	
French–English			
1 1345	0.0304	0.1025	Buchenwald concentration camp, Buchenwald, August 24
2 1623	0.3130	0.2960	Resistance movement, Poland
3 3005	0.0351	0.2249	Death marches, Death marches Holocaust, Schutzstaffel SS
4 3007	0.0113	0.0088	Europe, War
5 3009	0.1488	0.1247	War
6 3022	0.0568	0.0558	War, Country, Culture
7 3024	0.0010	0.0003	War
8 3025	0.0670	0.0401	War, Europe
9 3033	0.0975	0.0888	Palestine, Palestine region
German–English			
1 1133	0.1057	0.1044	Varian Fry, History, Marseille, Marseilles
2 1173	0.0461	0.0321	Art
3 3005	0.2131	0.1868	Schutzstaffel SS, Concentration camp, Concentration camps, Internment
4 3007	0.0058	0.0049	Europe
5 3009	0.1495	0.1256	War
6 3010	0.0002	0.0000	Germany, Property, Forced labor, Forced labour
7 3012	0.0003	0.0002	Germany, Jew, Jewish, Jewish People, Jews
8 3015	0.0843	0.0700	Jew, Jewish, Jewish People, Jews
9 3022	0.0658	0.0394	War, Holocaust, The Holocaust, Culture
10 3023	0.0100	0.0082	Holocaust, The Holocaust, Germany
11 3024	0.0006	0.0002	War, Holocaust, The Holocaust
12 3025	0.0857	0.0502	War, Jew, Jewish, Jewish People, Jews, Europe
13 3026	0.0021	0.0016	Concentration camp, Concentration camps, Internment
Spanish–English			
1 1173	0.0184	0.0077	Art, Literature
2 1345	0.0689	0.0596	Buchenwald concentration camp, Buchenwald, Allied powers, Allies of World War II, August 24
3 1624	0.0034	0.0003	Polonia, Poland, Holocaust, The Holocaust
4 3005	0.0685	0.0341	Posthumously, Schutzstaffel SS, Allied powers, Allies, Allies of World War II
5 3007	0.0395	0.0213	Europe, War
6 3009	0.1495	0.1256	War
7 3011	0.0413	0.0283	Holocaust, The Holocaust
8 3022	0.0661	0.0449	Holocaust, The Holocaust, Country, Culture
9 3024	0.0029	0.0016	Violence, War, Holocaust, The Holocaust
10 3025	0.0548	0.0371	War
11 3026	0.0036	0.0024	Partisans, War

Table 2). This produced improvements of 7% for the mono-lingual run, but only provided improvements of 5%, 1%, and 5% in French–English, Spanish–English, and German–English CL-SR effectiveness. The MAP value of the French–English (BabelFish+LEX+PRF) run provides the best results among all runs submitted by participants in this task.

6 Conclusions

Our experiments for the CLEF 2007 CL-SR task focused on the combination of standard MT with domain-specific translation resources. Our results indicate that combining domain-specific translation derived from Wikipedia with the output of standard MT can produce substantial improvements in CLIR retrieval effectiveness. For example, the French term ‘Hassidisme’ (in Query 1166) is translated to ‘Hasidic Judaism’ in English, ‘Varian Fry’ (in Query 1133) and ‘Marches de la mort’ (in Query 3005) are correctly detected as phrases and thus translated as ‘Varian Fry’ and ‘death marches’ in English, respectively. Further improvements can also be observed when combined with PRF. However, these trends are not observed consistently, and further investigations will focus on understanding differences in behaviour, and refining our procedures for training domain-specific translation resources.

Acknowledgement. Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST - 033104. The authors are solely responsible for the content of this paper.

References

1. Jones, G.J.F., Zhang, K., Lam-Adesina, A.M.: Dublin city university at clef 2006: Cross-language speech retrieval (cl-sr) experiments. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 794–802. Springer, Heidelberg (2007)
2. Robertson, S.E., Zaragoza, H., Taylor, M.: Simple BM25 Extension to Multiple Weighted Fields. In: Proceedings of the 13th ACM CIKM, pp. 42–49 (2004)
3. 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, pp. 792–799. Springer, Heidelberg (2006)
4. 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, pp. 744–759. Springer, Heidelberg (2006)

5. Oard, D.W., Wang, 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 744–758. Springer, Heidelberg (2007)
6. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference, pp. 109–126 (1994)
7. Porter, M.F.: An algorithm for suffix stripping. *Automated Library and Information Systems* 14(3), 130–137 (1980)
8. Beaulieu, M.M., Gatford, M.: Interactive Okapi at TREC-6. In: Proceedings of the 6th Text REtrieval Conference, pp. 143–168 (1997)
9. Adafre, S.F., de Rijke, M.: Discovering missing links in Wikipedia. In: Proceedings of the 3rd International Workshop on Link Discovery, Chicago, Illinois, pp. 90–97. ACM Press, New York (2005)
10. Declerck, T., Pèrez, A.G., Vela, O., Gantner, Z., Manzano-Macho, D.: Multilingual Lexical Semantic Resources for Ontology Translation. In: Proceedings of LREC, Genoa, Italy (2006)

What Can and Cannot Be Found in Czech Spontaneous Speech Using Document-Oriented IR Methods — UWB at CLEF 2007 CL-SR Track[★]

Pavel Ircing, Josef Psutka, and Jan Vavruška

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{ircing,psutka,sandokan}@kky.zcu.cz

Abstract. The paper presents an overview of the system build and experiments performed for the CLEF 2007 CL-SR track by the University of West Bohemia. We have concentrated on the monolingual experiments using the Czech collection only. The approach that was successfully employed by our team in the last year's campaign (simple tf.idf model with blind relevance feedback, accompanied with solid linguistic preprocessing) was used again but the set of performed experiments was broadened and a more detailed analysis of the results is provided.

1 Introduction

The Czech subtask of the CL-SR track, which was first introduced at CLEF 2006 campaign, is enormously challenging — let us repeat once again that the goal is to identify appropriate replay points (that is, the moments where the discussion about the queried topics starts) in a continuous stream of text generated by automatic transcription of spontaneous speech. Therefore, it is neither the standard document retrieval task (as there are no true documents defined) nor the fully-fledged speech retrieval (since the participants do not have the speech data nor the lattices, so they can't explore alternative hypotheses and must rely on one-best transcription). However, in order to lower the barrier of entry for teams proficient at classic document retrieval (or, for that matter, even total IR beginners), the last year's organizers prepared a so called Quickstart collection with artificially defined “documents” that were created by sliding 3-minute window over the stream of transcriptions with a 2-minute step (i.e., the consecutive documents have a one minute overlap).¹ The last year's Quickstart

[★] This work was supported by the Ministry of Education of the Czech Republic project No. LC536 and the Grant Agency of the Czech Academy of Sciences project No. 1QS101470516.

¹ It turned out later that the actual timing was different due to some faulty assumptions during the Quickstart collection design, but since the principle of the document creation remains the same, we will still use the “intended” time figures instead of the actual ones, just for the sake of readability.

collection was further equipped with both manually and automatically generated keywords (see [1] for details) but they have been shown to be of no benefit for IR performance [2] (the former for the timing problems, the latter for the problems with their assignment that yet remain to be identified) and thus have been dropped from this year's data. The scripts for generating such a quickstart collection with variable window and overlap times were also included in the data release.

2 System Description

Our current system largely builds upon the one that was successful in the last year's campaign [2], with only minor modifications and larger set of tested settings.

2.1 Linguistic Preprocessing

Stemming (or lemmatization) is considered to be vital for good IR performance even in the case of weakly inflected languages such as English; thus it is probably even more crucial for Czech as the representative of the richly inflectional language family. This assumption was experimentally proven by our group in the last year's CLEF CL-SR track [2]. Thus we have used the same method of linguistic preprocessing, that is, the serial combination of Czech morphological analyser and tagger [3], which provides both the lemma and stem for each input word form, together with a detailed morphological tag. This tag (namely it's first position) is used for stop-word removal — we removed from indexing all the words that were tagged as prepositions, conjunctions, particles and interjections.

2.2 Retrieval

All our retrieval experiments were performed using the Lemur toolkit [4], which offers a variety of retrieval models. 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. There are many different formulas for computation of the weights $w_{i,j}$, we have tested two of them, varying in the *tf* component:

Raw Term Frequency

$$w_{i,j} = tf_{i,j} \cdot \log \frac{d}{df_j} \quad (1)$$

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 .

BM25 Term Frequency

$$w_{i,j} = \frac{k_1 \cdot tf_{i,j}}{tf_{i,j} + k_1(1 - b + b \frac{l_d}{l_C})} \cdot \log \frac{d}{df_j} \quad (2)$$

where $tf_{i,j}$, d and df_j have the same meaning as in (1), l_d denotes the length of the document, l_C the average length of a document in the collection and finally k_1 and b are the parameters to be set.

The tf components for queries are defined analogously. The values of k_1 and b were set according to the suggestions made by [5] and [6], that is $k_1 = 1.2$ and $b = 0.75$ for computing document weights and $k_1 = 1$ and $b = 0$ for query weights².

We have also tested the influence of the blind relevance feedback. The simplified version of the Rocchio's relevance feedback implemented in Lemur [5] was used for this purposes. The original Rocchio's algorithm is defined by the formula

$$\mathbf{q}_{new} = \mathbf{q}_{old} + \alpha \cdot \mathbf{d}_R - \beta \cdot \mathbf{d}_{\bar{R}}$$

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 considered to be 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

3.1 Evaluated Runs

We have created 3 different indices from the collection — using original data and their lemmatized and stemmed version. There were 29 training topics and 42 evaluation topics defined by the organizers. We have first run the set of experiments for the training topics (see Table 1), comparing:

- Results obtained for the queries constructed by concatenating the tokens (either words, lemmas or stems) from the `<title>` and `<desc>` fields of the topics (TD - upper section of the table) with results for queries made from all three topic fields, i.e. `<title>`, `<desc>` and `<narr>` (TDN - lower section).
- Results achieved on the “original” Quickstart collection (i.e. 3-minute window with 1-minute overlap - Segments 3-1) with results computed using the collection created by using 2-minute window with 1-minute overlap (Segments 2-1).

² Setting $b = 0$ was actually not a choice, as this value is hard-set for queries.

In all cases the performance of raw term frequency (Raw TF) and BM25 term frequency (BM25 TF) is tested, both with (BRF) and without (no_FB) application of the blind relevance feedback. The mean Generalized Average Precision (mGAP) is used as the evaluation metric — the details about this measure can be found in [7].

Table 1. Mean GAP of the individual runs - training topics

		Segments 3-1				Segments 2-1			
		Raw TF		BM25 TF		Raw TF		BM25 TF	
		no_FB	BRF	no_FB	BRF	no_FB	BRF	no_FB	BRF
TD	words	0.0179	0.0187	0.0143	0.0172	0.0195	0.0233	0.0144	0.0174
	lemmas	0.0315	0.0358	0.0293	0.0337	0.0353	0.0458	0.0297	0.0364
	stems	0.0321	0.0364	0.0271	0.0343	0.0390	0.0463	0.0310	0.0377
TDN	words	0.0190	0.0209	0.0134	0.0169	0.0203	0.0219	0.0160	0.0196
	lemmas	0.0385	0.0435	0.0236	0.0343	0.0456	0.0536	0.0294	0.0396
	stems	0.0387	0.0414	0.0254	0.0365	0.0463	0.0510	0.0305	0.0401

Then we identified the 5 most promising/illustrative runs from the Table 1, repeated them for the evaluation topics and sent to the organizers for judgment. After receiving the relevance judgments for evaluation topics, we have replicated all the runs for those topics too (Table 2).

Table 2. Mean GAP of the individual runs - evaluation topics. Runs typeset in italics were submitted for official scoring.

		Segments 3-1				Segments 2-1			
		Raw TF		BM25 TF		Raw TF		BM25 TF	
		no_FB	BRF	no_FB	BRF	no_FB	BRF	no_FB	BRF
TD	words	0.0111	0.0128	0.0094	0.0126	0.0129	<i>0.0132</i>	0.0104	0.0113
	lemmas	0.0181	0.0208	0.0135	<i>0.0134</i>	0.0195	0.0217	0.0161	0.0144
	stems	0.0205	0.0223	0.0144	0.0173	0.0204	<i>0.0229</i>	0.0169	0.0198
TDN	words	0.0121	0.0154	0.0093	0.0113	0.0146	0.0171	0.0106	0.0131
	lemmas	0.0217	<i>0.0241</i>	0.0118	0.0155	0.0224	<i>0.0274</i>	0.0180	0.0168
	stems	0.0225	0.0232	0.0097	0.0116	0.0231	0.0263	0.0129	0.0137

3.2 Analysis of the Results

It turns out that the structure of the results for different experimental settings is similar for both the training and evaluation topics - thus we could observe the following general trends:

- Two minute “documents” seem to perform better than the three minute ones — probably the three minute segmentation is too coarse.

- The simplest raw term frequency weighting scheme generally outperforms the more sophisticated BM25 — one possible explanation is that in a standard document retrieval setup the BM25 scheme profits mostly from its length normalization component that is completely unnecessary in our case (remember that our documents all have approximately identical length by design).

The fact that both stemming and lemmatization boost the performance by about the same margin was already observed in the last year’s experiments.

In order to facilitate a more detailed result analysis, we selected the best performing “segment” of runs for both training and evaluation topics (i.e., the runs performed with TDN on Segments 2-1, using raw term frequency weighting and applying the relevance feedback - see the bold columns in Tables 1 and 2) and plotted the GAP for individual topics (see Figures 1 and 2).

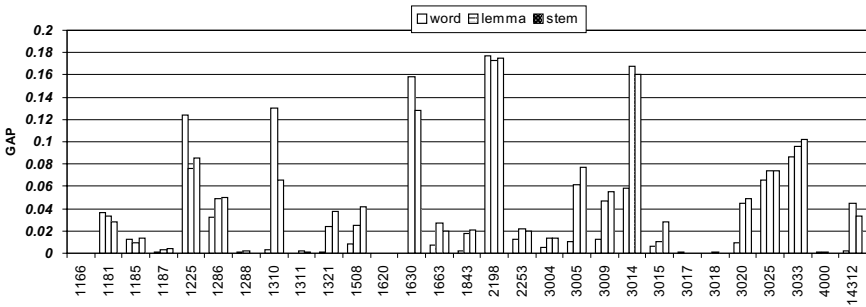


Fig. 1. GAP by topic for selected runs - training topics

Looking more closely at the results for individual topics, the reason for their failure/success given the employed retrieval approach is sometimes quite straightforward but in other cases rather nebulous. Let’s begin with the more apparent category. Unsuccessful topics often ask about rather abstract concepts without clear discriminative “keywords” (e.g. training topic 3017: “Social relations in concentration camp” or evaluation topic 3019: “Survival by previous professional identity”) — such queries are obviously not handled well with *tf.idf* model which essentially relies on matching terms from the query and terms from the documents. Similarly, other topics failed because even though they do contain discriminative terms, those terms are not found in the collection (e.g. training topic 1166: “Hasidism” or evaluation topic 1192: “Kindertransport”). Conversely, topics with highly discriminative keywords that are present also in the collection rank on the very top (training topics 2198: “Sonderkommando”, 3014: “Zionism” and 1630: “Eichmann witnesses”). On the other hand, the failure and, even more notably, the success of some topics remain mysterious — most flagrant

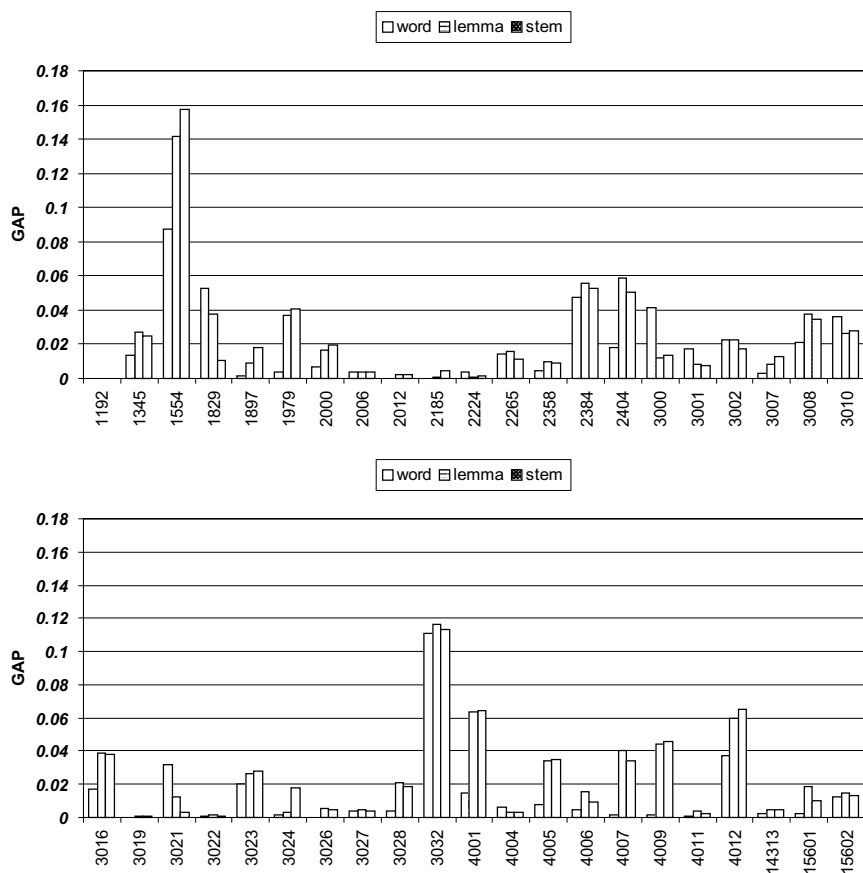


Fig. 2. GAP by topic for selected runs - evaluation topics

example being the evaluation topic 3032 which is a clear case of an “abstract” topic (“Survivor guilt”), yet it has the second-best score.

4 Conclusion

In the CLEF 2007 CL-SR task, we have made just a little step further towards successful searching of Czech spontaneous speech and probably also reached the performance limits of the passage retrieval approach to speech searching. In order to make a bigger progress, we would need to really take the speech part of the task into account — that is, to use the speech recognizer lattices when searching for the desired information, or even to modify the ASR components so that it will be more likely to produce output useful for IR (for example, enrich the language model with rare named entities that are currently often being misrecognized).

References

1. 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.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
2. Ircing, P., Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
3. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech), Karolinum, Prague (2004)
4. Carnegie Mellon University and the University of Massachusetts: The Lemur Toolkit for Language Modeling and Information Retrieval (2006), <http://www.lemurproject.org/>
5. Zhai, C.: Notes on the Lemur TFIDF model. Note with Lemur 1.9 documentation, School of CS, CMU (2001)
6. Robertson, S., Walker, S.: Okapi/Keenbow at TREC-8. In: The Eight Text REtrieval Conference (TREC-8) (1999)
7. 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)

Using Information Gain to Filter Information in CLEF CL-SR Track

M.C. Díaz-Galiano, M.T. Martín-Valdivia, M.A. García-Cumbreras,
and L.A. Ureña-López

SINAI Research Group, Computer Science Department, University of Jaén, Spain
`{mcdiaz,maite,magc,laurena}@ujaen.es`

Abstract. This paper describes the first participation of the SINAI team in the CLEF 2007 CL-SR track. The SINAI team has only participated in the English task. The English collection includes segments of audio speech recognition and topics to evaluate the information retrieval systems. This collection contains interviews with survivors of the Holocaust manually segmented. Moreover, each segment includes different fields with extra information. The topics to evaluate the English task are available in Czech, English, French, German, Dutch and Spanish. This year, the team only wants to establish a first contact with the task and the collection. Thus, the collection has been pre-processed using the Information Gain technique in order to filter the fields with most relevant information. The Lemur toolkit has been the Information Retrieval system used in the experiments.

1 Introduction

This paper presents the first participation of the SINAI research group at the CLEF CL-SR track. Our main goal was to study the use of the Information Gain technique over a collection of transcribed texts. We have already used this measure in order to filter the fields of a collection with metadata [1].

Information Gain (IG) is a measure that allows to select the meta-data that contribute more information to the system, ignoring those that not only provide zero information but which, at times, can even introduce noise, thus distorting the system response. Therefore, it is a good candidate for selecting the meta-data that can be useful for the domain in which the collection is used. Information Gain has been used in numerous studies [2], most of them centered on classification. Some examples could be Text Categorization [3], Machine Learning [4] or Anomaly Detection [5].

The CLEF CL-SR track has two tasks [6], namely, the English task and the Czech task. We only have participated in the former. The English collection includes 8,104 segments of audio speech recognition and 105 topics. To create this collection, interviews with survivors of the Holocaust were manually segmented to form topically coherent segments by subject matter experts at the Survivors of the Shoah Visual History Foundation. All the topics for the English task are available in Czech, English, French, German, Dutch and Spanish. These 105

topics consist on 63 training topics from 2006, 33 test topics from 2006 and 9 new topics for which relevance data is not currently available. These 9 topics have been included in order to support possible future construction of new relevance assessment pools. Therefore, the results are reported only for the 33 test topics.

IG is usually used for feature set selection so this work treats the different fields in the document as feature sets. The following section describes the field selection process with Information Gain. In Section 3, we explain the experiments and obtained results. Finally, conclusions are presented in Section 4.

2 Field Selection with Information Gain

We have used the Information Gain measure [7] to select the best XML fields in the collection.

The method applied consists of computing the Information Gain for each field in the collection. Let C be the set of cases and E the value set for the E field. 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 field,

$H(C)$ is the entropy of the set of cases C

$H(C|E)$ is the relative entropy of the set of cases C conditioned by the E field

Both, $H(C)$ and $H(C|E)$ are calculated based on the frequencies of occurrence of the fields 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 field 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 each field in the collection, its Information Gain is computed. Then, the fields selected to compose the final collection are those showing higher values of Information Gain. Once the document collection was generated, experiments were conducted with the Lemur¹ retrieval information system.

3 Experiment Description and Results

Our main goal was to study the effectiveness of filtering fields using Information Gain in the text collection. For that purpose, we have accomplished several experiments using all the fields in the collection to identify the best field percentage

¹ <http://www.lemurproject.org/>

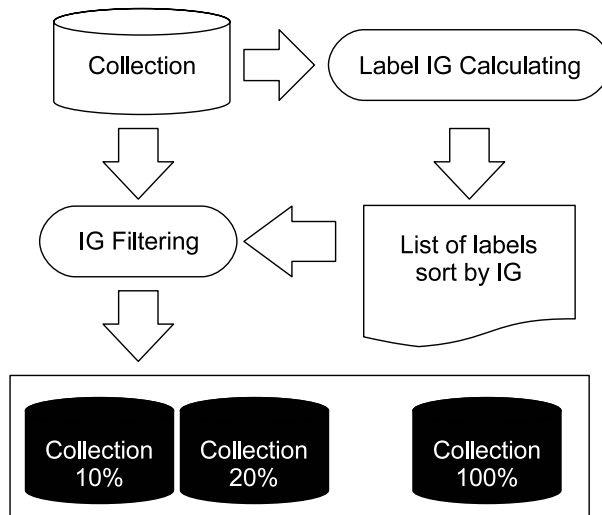


Fig. 1. Field selection using Information Gain filtering

with experiments preserving 10%, 20%...100% of fields (Figure 1). It is important to note that rare values of a field lead to very high Information Gain values as for the DOCNO field, whose values are unique for each document. This is the expected behavior for Information Gain, because by knowing the DOCNO field we could retrieve the exact document. Unfortunately, this field is useless, since we expect to retrieve documents based on the words of these documents and the DOCNO field is not a valid word. For this reason we calculate a new value based on document frequency (DF). The global document frequency average (GDFA) value of a field was calculated this way:

- We calculate the DF of each word of a field in a document.
- Then we obtain the DF average of this field, calculating the sum of the DF of each word and dividing it by the number of words of this field in this document.
- Finally, we calculate the GDFA of a field in the collection, using the sum of all the DF average of this field in each document and dividing it by the number of document where the field exists.

The fields with a GDFA less as 4 are put in the bottom of the list (Figure 2). Table 1 shows the Information Gain values of the collection fields sorted by Information Gain and applying GDFA sorting.

Therefore, we have run ten experiments (with ten Information Gain collections) for each list of topics in English, Dutch, French, German and Spanish. However, we have only sent five runs, since the organizers limited the number of submissions. These five runs were:

- SinaiEn050: English queries and collection with 50% of fields.
- SinaiEn100: English queries and collection with 100% of fields.

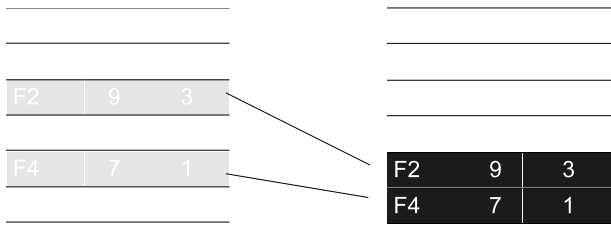


Fig. 2. Example of field sorted by IG and GDFA

Table 1. List of fields sorted by Information Gain (IG)

Fields	IG	GDFA	Fields Percent
DOC/SUMMARY	12.9834	2012.07	10%
DOC/ASRTEXT2004A	12.9792	1918.77	20%
DOC/ASRTEXT2006B	12.9775	1935.68	30%
DOC/AUTOKEYWORD2004A2	12.9574	4463.32	40%
DOC/AUTOKEYWORD2004A1	12.9521	3484.73	50%
DOC/ASRTEXT2006A	12.6676	1770.60	50%
DOC/MANUALKEYWORD	12.6091	3355.97	60%
DOC/ASRTEXT2003A	12.5953	1665.31	70%
DOC/NAME	11.9277	46.43	80%
DOC/INTERVIEWDATA	8.4755	239.81	90%
DOC/DOCNO	12.9844	1.00	100%

- SinaiFr100: English queries translated from Frech and collection with 100% of fields.
- SinaiSp050: English queries translated from Spanish and collection with 50% of fields.
- SinaiSp100: English queries translated from Spanish and collection with 100% of fields.

French, German and Spanish topics have been translated to English using a translation module.

The translation module used is SINTRAM (SINai TRAnslation Module), our Machine Translation system, which uses some online Machine Translators for each language pair and implements some heuristics to combine the different translations [8]. After a complete research, we have found that the best translators were:

- Systran for Dutch, French and German
- Prompt for Spanish

The experiments have been carried out with Lemur Information Retrieval system. After a complete experimentation with several weighting functions and

Table 2. MAP values for all experiments

Field Percent	Dutch	English	French	German	Spanish
10%	0,0790	0,0925	0,0925	0,0508	0,0982
20%	0,0680	0,0662	0,0662	0,0449	0,0773
30%	0,0607	0,0619	0,0619	0,0404	0,0616
40%	0,0579	0,0569	0,0569	0,0408	0,0628
50%	0,0560	0,0515	0,0515	0,0391	0,0579
60%	0,0643	0,0609	0,0609	0,0493	0,0741
70%	0,0623	0,0601	0,0601	0,0474	0,0735
80%	0,0622	0,0597	0,0597	0,0473	0,0735
90%	0,0621	0,0601	0,0601	0,0470	0,0737
100%	0,0619	0,0597	0,0597	0,0470	0,0737

the use or not of Pseudo-Relevance Feedback (PRF), the best configuration was KL-divergence with PRF.

Table 2 shows the results for all the experiments. The experiments with Spanish and Dutch queries translations are better than the other ones (in experiments from 20% of fields to 100%).

Table 2 shows that the use of some fields does not improve the system results. The SUMMARY and MANUALKEYWORD fields have been manually created by a human expert and are the best ones (included in 10% and 60% respectively). Those fields obtain the best results in the retrieval system. In table 2, the MAP value of experiments from 100% of fields to 50% decreases in the all languages, but in the experiment with 60% of fields, the MAP value increases also in the all languages. Table 1 shows that the field included at the 60% is the MANUALKEYWORD field. These results indicate that the automatic fields (ASRTEXT2003A, ASRTEXT2004A, ASRTEXT2006A, ASRTEXT2006B, AUTOKEYWORD2004A1, AUTOKEYWORD2004A2) are not the best ones.

Results of the cross-lingual experiments show that Spanish and Dutch translations are better than the other experiments. The Spanish experiments confirm the good results obtained in the ImageCLEF ad-hoc task [9] this year.

4 Conclusions

In our first participation in CLEF CL-SR we have used Information Gain in order to find the best fields in the collection. The IG values of each field are very similar, so that it is very difficult to select the best ones. Moreover, the corpus does not have many fields, and one of them (SUMMARY field) obtains quite good results in our experiments. In other experiments with a similar collection [10] the use of field SUMMARY and MANUALKEYWORD obtains very good results too. Therefore, the IG strategy cannot find the best fields in the CL-SR collection.

Our next step will focus on improving the query expansion using an ontology [11]. This approach has obtained the best results for us in other tasks of CLEF [9].

Acknowledgements

This project has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id.514 granted by the University of Jaén.

References

1. Ureña-López, L.A., Díaz-Galiano, M.C., Montejo-Raez, A., Martín-Valdivia, M.T.: The Multimodal Nature of the Web: New Trends in Information Access. UP-GRADE (The European Journal for the Informatics Professional). Monograph: Next Generation Web Search, 27–33 (2007)
2. Quinlan, J.R.: Induction of Decision Trees Machine Learning (1), 81–106 (1986)
3. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of ICML 1997, 14th International Conference on Machine Learning (1997)
4. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1996)
5. Lee, W., Xiang, D.: Information-Theoretic Measures for Anomaly Detection. In: Proc. of the 2001 IEEE Symposium on Security and Privacy (2001)
6. Oard, D.W., Wang, 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley-Interscience, Chichester (2006)
8. García-Cumbreras, M.A., Ureña-López, L.A., Martínez-Santiago, F., Perea-Ortega, J.M.: BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2006) (2006)
9. Díaz-Galiano, M.C., García-Cumbreras, M.A., Martín-Valdivia, M.T., Montejo-Raez, A., Ureña-López, L.A.: SINAI at ImageCLEF 2007. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)
10. Alzghool, M., Inkpen, D.: University of Ottawa's Participation in the CL-SR Task at CLEF 2006. In: Working Notes of the 2006 CLEF Workshop, Alicante, Spain (September 2006)
11. Alzghool, M., Inkpen, D.: Fusion Experiments for the Cross Language Speech Retrieval Task at CLEF 2007. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)

Overview of WebCLEF 2007

Valentin Jijkoun and Maarten de Rijke

ISLA, University of Amsterdam
{jijkoun,mdr}@science.uva.nl

Abstract. This paper describes the WebCLEF 2007 task. The task definition—which goes beyond traditional navigational queries and is concerned with undirected information search goals—combines insights gained at previous editions of WebCLEF and of the WiQA pilot that was run at CLEF 2006. We detail the task, the assessment procedure and the results achieved by the participants.

The WebCLEF 2007 task combines insights gained from previous editions of WebCLEF 2005–2006 [1, 6] and the WiQA 2006 pilot [3, 4], and goes beyond the navigational queries considered at WebCLEF 2005 and 2006. At WebCLEF 2007 we consider so-called undirected informational search goals [5] in a web setting: “I want to learn anything/everything about my topic.” A query for topic X might be interpreted as “Tell me about X.”

In the remainder of the paper we detail the task, the assessment procedure and the results achieved by the participants.

1 Task Description

As key starting points for defining the WebCLEF task we took several issues into account. First, the task should correspond as close as possible to some real-world information need with a clear definition of the use case. Second, multi- and cross-linguality should be natural (or even essential) for the task in the CLEF setting. Next, the collection(s) used in the task should be a natural source of choice for the user’s information need. Then, collections, topics and assessors’ judgements, resulting from the task should be re-usable in future. Finally, the task should be challenging for the state-of-the-art IR and NLP technology.

1.1 Task Model

Our hypothetical user is a knowledgeable person, perhaps even an expert, writing a survey article on a specific topic with a clear goal and audience, for example, a Wikipedia article, or a state of the art survey, or an article in a scientific journal. She needs to locate items of information to be included in the article and wants to use an automatic system to help with this. The user does not have immediate access to offline libraries and only uses online sources.

The user formulates her information need (the topic) by specifying:

- a short *topic title* (e.g., the title of the survey article),
- a free text *description* of the goals and the intended audience of the article,
- a list of *languages* in which the user is willing to accept the found information,
- an optional list of *known sources*: online resources (URLs of web pages) that the user considers to be relevant to the topic and information from which might already have been included in the survey article, and
- an optional list of *Google retrieval queries* that can be used to locate the relevant information; the queries may use site restrictions (see examples below) to express the user’s preferences.

Here’s an example of a WebCLEF 2007 topic describing an information need:

- topic title: *Significance testing*
- description: I want to write a survey (about 10 screen pages) for undergraduate students on statistical significance testing, with an overview of the ideas, common methods and critiques. I will assume some basic knowledge of statistics.
- language(s): English
- known source(s): http://en.wikipedia.org/wiki/Statistical_hypothesis_testing ; http://en.wikipedia.org/wiki/Statistical_significance
- retrieval queries: significance testing; site:mathworld.wolfram.com significance testing; significance testing pdf; significance testing site:en.wikipedia.org

Defined in this way, the task model corresponds to addressing undirected informational search goals, that are reported to account for over 23% of web queries [5].

Each participating team was asked to develop several topics and subsequently assess responses of all participating systems for the created topics and.

1.2 Data Collection

In order to keep the idealized task as close as possible to the real-world scenario (when typically there are many documents somehow relevant to the user’s information need) but still tractable (i.e., keeping the size of the collection is manageable), our collection is defined per topic. Specifically, for each topic, the subcollection for the topic consists of the following set of documents along with their URLs:

- all “known” sources specified for the topic;
- the top 1000 (or less, depending at the actual availability at the time of querying) hits from Google for each of the retrieval queries specified in the topic, or for the topic title if the queries are not specified;
- for each online document included in the collection, its URL, the original content retrieved from the URL and the plain text version of the content are provided. The plain text conversion is only available for HTML, PDF and

Postscript documents. For each document, the subcollection also specifies its origin: which query or queries were used to locate it and at which rank(s) in the Google result list it was found.

1.3 System Response

For each topic description, a response of an automatic system consists of a ranked list of plain text snippets (of arbitrary length) extracted from the sub-collection of the topic. Each snippet should indicate what document of the sub-collection it comes from.

2 Assessment

In order to comply with the task model, the manual assessment of the responses of the systems was done by the topic creators. The assessment procedure was somewhat similar to assessing answers to OTHER questions at TREC 2006 Question Answering task [8].

The assessment was blind. For a given topic, all responses of all system were pooled into anonymized sequence of text segments (snippets). To limit the amount of required assessments, for each topic only first 7,000 characters of each response were included (according to the ranking of the snippets in the response). The cut-off point 7,000 was chosen so that for at least half of the submitted runs the length of the responses was at least 7,000 for all topics. For the pool created in this way for each topic, the assessor was asked to make a list of nuggets, atomic facts, that, according to the assessor, should be included in the article for the topic. A nugget may be linked to character spans in the responses, so that all spans linked to one nugget express this atomic fact. Different character spans in one snippet in the response may be linked to more than one nugget. The assessors used a GUI to mark character spans in the responses and link each span to the nugget it expresses (if any). Assessors could also mark character spans as “known” if they expressed fact relevant for the topic but already present in one of the known sources.

Similar to INEX [2] and to some tasks at TREC (i.e., the 2006 Expert Finding task [7]) assessment was carried out by the topic developer, i.e., by the participants themselves.

Table 1 gives the statistics for the 30 test topics and for the assessments of the topics.¹

3 Evaluation Measures

The evaluation measures for the task are based on standard precision and recall. For a given response R (a ranked list of text snippets) of a system S for a topic T we define:

¹ Full definition of the test topics is available from
<http://ilps.science.uva.nl/WebCLEF/WebCLEF2007/Topics>.

Table 1. Statistics for WebCLEF 2007 topics and assessments. For each topic we show: topic id and title, accepted languages, the number of “known” sources (web pages), the total number of snippets assessed from all submissions, the total number of spans marked as “known”, the total number of spans attached to one of the nuggets, and the total length (the number of characters) in these spans.

Id	Topic title	Langs	known		total	known		chars
			srcs	snippets		spans	marked	
1	Big Bang Theory	en	3	258	2	164	36591	
2	Symptoms Avian Influenza or bird flu	en	2	384	0	46	12595	
3	The architecture of Borobudur Temple	en	2	249	0	29	12198	
4	Sistemas de calefacción por Biomasa	es,en,pt,it	6	324	2	5	3739	
5	Sistemas biométricos de autentificación	es,en,pt,it	6	241	7	17	4714	
6	revistas científicas open access	es,en,pt,it	5	341	4	13	1586	
7	Magnum Opus	en	1	308	3	3	765	
8	Bloomsday (Band)	en	1	261	6	4	596	
9	Belayneh Densamo	en	1	412	16	1	197	
10	The Empire in Africa	en	2	235	3	25	6402	
11	Gaelic Games	en	1	261	17	11	2706	
12	schatten van voorwaardelijke kansen	nl	1	291	14	0	0	
13	sentiment analysis for European languages (non-English)	nl,en	1	254	4	2	450	
14	European blog search engines	nl,en	1	273	0	6	497	
15	verhuistips	nl	2	238	26	4	948	
16	Yasujiro Ozu	nl,en	3	268	10	10	4570	
17	Visa regulations regarding HIV status of applicants	en,nl	0	269	0	31	6237	
18	Holidays on Maldives	en	0	281	0	21	1798	
19	Comparison of retrieval models in Information Retrieval	en	2	238	0	29	5915	
20	ATM (automated teller machine) fraud	en,nl	1	264	0	72	13158	
21	iPhone opinions	en,nl	0	290	0	35	4388	
22	school education in The Netherlands	en,nl	1	251	9	16	4476	
23	Details on obtaining Russian tourist visa for foreigners	en,nl	0	285	0	39	7553	
24	Albrecht Drer's "Passions" engravings and woodcuts	en	1	387	0	6	807	
25	Human-induced climate change: pro and cons	en,nl	0	260	0	27	7345	
26	Plastic tableware and the environment	en,nl	0	275	0	26	3887	
27	Details on interpretation of Ravel's "Gaspard de la Nuit"	en	1	250	4	21	8024	
28	Nabokov's "Invitation to a Beheading"	en	1	258	11	20	2702	
29	Durability of digital storage media	en	0	279	0	9	1605	
30	Madonna's books for children	en	1	253	0	45	9019	

- *recall* as the sum of character lengths of all annotated spans in R linked to nuggets, divided by the total sum of annotated span lengths in the responses for T in all submitted runs.
- *precision* as the number of characters that belong to at least one annotated span, divided by the total character length of the system's response.

Note that the evaluation measures described above differ slightly from the measures originally proposed in the task description.² The original measures were based on the fact that spans are linked to nuggets by assessors: as described in section 2, different spans linked to one nugget are assumed to bear approximately the same factual content. Then, in addition to character-based measures above, a *nugget-based recall* can be defined based on the number of nuggets (rather than lengths of character spans) found by a system. However, an analysis of the assessments showed that some assessors used nuggets in a way not intended by the assessment guidelines: namely, to group related rather than synonymous character spans. We believe that this misinterpretation of the assessment guidelines indicates that the guidelines are overly complicated and need to be simplified in future edition of the task. As a consequence, we did not use nugget-based measures for evaluation.

4 Runs

In total, 12 runs were submitted from 4 research groups. To provide a baseline for the task, we created an artificial run: for each topic, a response of the baseline was created as the ranked list of at most 1000 snippets provided by Google in response to retrieval queries from the topic definition. Note that the Google web search engine was designed for a task very different from WebCLEF 2007 (namely, for the task of web page finding), and therefore the evaluation results of our baseline can in no way be interpreted as an indication of Google's performance.

Table 5 shows the submitted runs with the basic statistics: the average length (the number of bytes) of the snippets in the run, the average number of snippets in the response for one topic, and the average total length of response per topic.

5 Results

Table 5 shows the evaluation results for the baseline and the submitted runs: precision and recall at three different cut-off points. Since the sizes of the submitted runs varied substantially (Table 5), the cut-off points were chosen to enable comparison across runs.

Table 5 indicates that most runs outperform the baseline, or show a similar performance. Two of the runs (*UVA par vs* and *UVA par wo*) show the best

² See <http://ilps.science.uva.nl/WebCLEF/WebCLEF2007/Tasks>.

Table 2. Simple statistics for the baseline (Google snippets) and the 12 submitted runs

Participant	Run	Average snippet length	Average snippets per topic	Average response length per topic
Baseline	Google snippets	145	898	131041
School of Computing, Dublin City University	DCU run1 simple	118	30	3552
	DCU run2 parsed	137	27	3770
	DCU run2 topfilter	112	29	3346
Faculty of Computer Science, University of Indonesia	UIWC07odwgstr	151	10	1522
	UIWC07uw10	155	10	1559
	UIWC07wstr	152	10	1530
REINA Research Group, University of Salamanca	USAL reina0.25	833	50	41680
	USAL reina0	832	50	41658
	USAL reina1	833	50	41708
ISLA, University of Amsterdam	UVA par vs	254	29	7420
	UVA par wo	277	25	7158
	UVA sent wo	214	33	7225

Table 3. Evaluation results for the baseline (Google snippets) and the 12 submitted runs calculated at cut-off points 1,500, 3,500 and 7,000 bytes for a response for one topic

Run	@ 1,500 bytes		@ 3,500 bytes		@ 7,000 bytes	
	P	R	P	R	P	R
Google snippets	0.13	0.3	0.11	0.07	0.08	0.11
DCU run1 simple	0.07	0.02	0.08	0.05	—	—
DCU run2 parsed	0.10	0.03	0.10	0.06	—	—
DCU run2 topfilter	0.08	0.02	0.08	0.04	—	—
UIWC07odwgstr	0.11	0.03	—	—	—	—
UIWC07uw10	0.09	0.02	—	—	—	—
UIWC07wstr	0.11	0.03	—	—	—	—
USAL reina0.25	0.11	0.03	0.14	0.09	0.16	0.20
USAL reina0	0.11	0.03	0.13	0.08	0.14	0.18
USAL reina1	0.11	0.03	0.14	0.09	0.16	0.21
UVA par vs	0.19	0.05	0.20	0.13	0.20	0.26
UVA par wo	0.15	0.04	0.20	0.13	0.20	0.25
UVA sent wo	0.10	0.03	0.09	0.06	0.09	0.11

performance for all cut-off points. Two other runs (*USAL reina0.25* and *USAL reina1*) show a comparable performance.

One unexpected phenomenon is that for all runs (except the baseline) the precision grows as the cut-off point increases. This might indicate that although systems manage to find relevant information snippets in the collection, the ranking of the found snippets is far from optimal.

6 Conclusions

We described WebCLEF 2007. This was the first year in which a new task was being assessed, one aimed at undirected information search goals. While the number of participants was limited, we believe the track was a success, as most submitted runs outperformed the Google-based baseline. For the best runs, in top 7,000 bytes per topic about 1/5 of the text was found relevant and important by the assessors.

The WebCLEF 2007 evaluation also raised several important issues. The task definition did not specify the exact size of a system's response for a topic, which has made a comparison across systems problematic. Furthermore, assessor's guidelines appeared to be overly complicated: not all assessors used nuggets as was intended by the organizers. We will address these issues in the 2008 edition of the task.

Acknowledgments

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 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

1. Balog, K., Azzopardi, L., Kamps, J., de Rijke, M.: Overview of WebCLEF 2006. In: CLEF 2006 (2007)
2. Fuhr, N., Lalmas, M., Trotman, A. (eds.): Comparative Evaluation of XML Information Retrieval Systems: 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006. Springer, Heidelberg (2007)
3. Jijkoun, V., de Rijke, M.: Overview of the WiQA task at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
4. Jijkoun, V., de Rijke, M.: WiQA: Evaluating Multi-lingual Focused Access to Wikipedia. In: Sakai, T., Sanderson, M., Evans, D.K. (eds.) Proceedings EVIA 2007, pp. 54–61 (2007)
5. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: WWW 2004: Proceedings of the 13th intern. conf. on World Wide Web, pp. 13–19. ACM Press, New York (2004)
6. 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, pp. 810–824. Springer, Heidelberg (2006)
7. Soboroff, I., de Vries, A.P., Craswell, N.: Overview of the TREC 2006 Enterprise Track. In: The Fifteenth Text REtrieval Conference (TREC 2006) (2007)
8. Voorhees, E.M., Dang, H.T.: Overview of the TREC 2005 question answering track. In: The Fourteenth Text REtrieval Conference (TREC 2005) (2006)

Segmentation of Web Documents and Retrieval of Useful Passages

Carlos G. Figuerola, José L. Alonso Berrocal, and Angel F. Zazo Rodríguez

University of Salamanca, REINA Research Group
c/ Fco. de Vitoria, 6-16, 37008 Salamanca, Spain
reina@usal.es
<http://reina.usal.es>

Abstract. This year's WebCLEF task was to retrieve snippets and pieces from documents on various topics. The extraction and the choice of the most widely used snippets can be carried out using various methods. This article illustrates the segmentation process and the choice of snippets produced in this process. It also describes the tests carried out and their results.

1 Introduction

This year's task focused on the retrieval of text snippets or fragments of web pages containing information on a specific topic; additionally, these snippets had to belong to a specific group of predetermined languages. As a starting point, we had a description of each topic (along with a title and a broader description), as well as a set of documents and well-known sources relating to the specific subject. Moreover, for each topic we did at least one search on *Google* with the first n documents retrieved.

As far as we are concerned, the general focus was to consider all the documents for each topic retrieved through *Google* as a set of documents that we could work with. Since the task was to obtain snippets, these documents were broken down into smaller fragments, with each fragment being regarded as a separate document.

As the query, we can use the description that we have for each topic. Additionally, this query can be enriched with further terms, provided by the well-known sources for this topic. We can also use the available anchors that are indicated on the documents retrieved through *Google*.

Finally, we must install filters or restrictions which can rule out any documents retrieved that do not belong to the group of predetermined languages. In this way, the task can be tackled as a common retrieval problem and, as a result, be applied as a conventional technique.

2 Preparation of Web Document Collection

As mentioned before, for each topic, the document collection was formed by the snippets available from the documents retrieved through *Google*. For each

of these topics, one or more searches were carried out on *Google* and for each of these searches the first n documents retrieved were taken, yielding a variable number of documents per topic.

All searches on *Google* for the same subject were considered as equal. Thus, for each document retrieved on *Google*, we had to elicit the original document, transform it into text, break it down into snippets, obtain the terms for each snippet and calculate their weights. The task organizers had already solved the first part of these operations by giving us the original documents and converting them into text.

In general, the conversion into plain text was good, a task that is not easy to accomplish. However, the encoding of characters was inconsistent, although it was stated that the plain text was encoded in UTF-8. For languages using characters not contained in the ASCII standard, the encoding and decoding of such characters is a source of problems. Just the detection of the coding system is, in many cases, problematic. For example, we used the Universal Encoding Library Detector (generally known as *chardet* [1]), a module for Python based on the libraries for encoding detection used by Mozilla. Chardet indicates that, surprisingly, the majority of the text versions were in Latin-2.

2.1 Division of Documents

Various techniques can be used to break down documents and to obtain more or less short text passages. In general, some are based on snippet size and can be estimated in bytes (or in characters) or in words. Others address the division of phrases and paragraphs [2]. The first type of technique obtains fragments that are more homogeneous in size but often lacking in meaning, since the starting point is blind.

The other techniques tend to produce snippets of very different sizes. Moreover, the implementation is not always easy; in many cases the conversion of HTML documents into plain text leaves out the space between paragraphs and the differences between soft and hard line breaks, and also eliminates or conceals structural elements, such as tables [3].

A simple approach like the choice of orthographic characters [4], such as the full stop, to break down a text tends to produce snippets that are too short and therefore, of little use in this type of task. In our case, we used a mixed approach. After several trials, we decided that a suitable size for each passage was 1500 characters, but since we were looking for fragments that made sense, our fragmentor would search for the chosen orthographic character (full stop) closest to the 1500 characters and split the text there.

2.2 Other Lexical Analysis Operations

Additionally, other procedures were carried out: conversion into lower case letters, omission of accents, omission of stop words (through a long list of stop words for all possible languages) and implementation of a single s-stemmer.

Each fragment obtained by these procedures was regarded as a separate document. From the documents obtained in this way the terms were extracted and

then weighted with the ATU weight scheme [5] and applied to the well-known vector model.

3 Building the Queries

Our objective was to complete the task by using conventional retrieval techniques, at least those already known. From the document collection formed with the snippets obtained, we had to select those that would be most useful for each topic. The key to our approach was to compose queries that could lead to an appropriate result. In order to compose these queries, we had several sources of information available. Firstly, we had topics with short titles and brief explanations, and for each topic we also had a number of documents termed as “known sources” with the complete texts. We also had the queries made on *Google* to obtain information on each topic. However, as the set of documents came from the results of the aforementioned queries, the information contained in such queries had already been used.

We were thus able to use the topics (the titles as well as the explanations) as a core or basis for each query, and enrich or increase them with terms available from the “known sources” [6]. The “known sources” are complete documents, some of them rather long, which can contain a lot of terms. We wondered whether this might bring too much noise to the queries. One possible solution is to weight the terms coming from the “known sources” in a different manner [7] to those coming from the title and description for each topic.

It is also possible to consider different structures or fields within the “known sources”, given that the majority are HTML pages (title, body, headings, meta tags, etc.) Previous experiments in earlier editions of CLEF demonstrate the importance of some of these fields for retrieval and the scarce importance of others [8]. The most interesting field, in this case, is the anchor of the back links. However, given that we have a rather small set of documents, we do not have many back links to work with. Nevertheless, those from the “known sources” which point to some of the documents retrieved from *Google* seem especially important.

Thus, in the queries we used the terms from the titles and descriptions of the topics, as well as the terms from the anchors mentioned previously. To this we added the terms from the “known sources” but weighted in a different manner. In previous editions of WebCLEF we worked on the use of different sources of information in retrieval, and on how to mix and merge these sources. On this occasion we chose to modify the weights of the terms, operating on their frequency in each document. The weighting scheme chosen for the queries was ATU (slope=0.2), which is why this weight is directly proportional to the frequency of the term in the document; we thus established a coefficient with which to multiply that frequency.

The runs carried out vary as a function of this coefficient: one of them maintains the original frequency, so the terms coming from the “known sources” are weighted the same as those of the topics. Another run weighted the terms of the “known sources” reducing the weight by one-fourth ($freq. \times 0.25$), and the third run did not use any terms at all from the “known sources”.

4 Results

The results of the three runs show little difference between them. It seems that using the terms from the “known sources” is more useful than not, but changing weights for the terms from the “known sources” produces little differences. In addition, we must take into account that some topics (nearly half of them) do not produce any useful results. We did not apply any restrictions or filters for the predetermined languages; however, there are a majority of web pages that are in english.

It is possible that some restrictions based on the type of information contained in the snippets from the retrieved documents could have been of interest. For example, several of these fragments are references to other sources of information (bibliographical references, academic courses or subjects on these topics, etc.) It seems that this type of information is not very useful for this task. Of another side, an important part of retrieved documents is repeated (the same document, but with different URL). A simple duplicated document control would have, probably, made arise other worse located in the ranking of retrieved, but perhaps useful.

Table 1. Official Runs and Results

	run 0	run 0.25	run 1
Precision	0.1415	0.1599	0.1624
Recall	0.1796	0.2030	0.2061

5 Conclusion

We have based our work on the construction of queries with terms coming from the “known sources”, along with the terms from the descriptions of the topics. The use of the terms from the “known sources” produces better results, although not in a dramatic way. However, the way in which the documents were broken down, as well as the selection of them based on language and on the type of information contents seems to have a strong impact on the results.

References

1. Pilgrim, M.: Universal encoding detector, <http://chardet.feedparser.org>
2. Zazo, Á.F., Figuerola, C.G., Alonso Berrocal, J.L., Rodríguez, E.: Reformulation of queries using similarity thesauri. *Information Processing & Management* 41(5), 1163–1173 (2005)
3. Yu, S., Cai, D., Wen, J.R., Ma, W.Y.: Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In: *Proceedings of the Twelfth International World Wide Web Conference, WWW 2003, Budapest, Hungary, 20-24 May 2003*, pp. 11–18. ACM, New York (2003)
4. Mikheev, A.: Tagging sentence boundaries. In: *Proceedings of the First Meeting of the North American Chapter of the Computational Linguistics (NAACL 2000)*, pp. 264–271. Morgan Kaufmann, San Francisco (2000)

5. 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 18–22, 1996, pp. 21–29. ACM, New York (1996) (Special Issue of the SIGIR Forum)
6. Lee, J.H.: Combining multiple evidence from different relevance feedback methods. Technical report, Center for Intelligent Information Retrieval (CIIR), Department of Computer Science, University of Massachusetts (1996)
7. 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)
8. Figuerola, C.G., Alonso Berrocal, J.L., Zazo Rodríguez, Á.F., Rodríguez, E.: REINA at WebCLEF 2006: Mixing fields to improve retrieval. In: Nardi, A., Peters, C., Vicedo, J. (eds.) *ABSTRACTS CLEF 2006 Workshop*, Alicante, Spain, 20–22 September. Results of the CLEF 2006 Cross-Language System Evaluation Campaign (2006)

Using Centrality to Rank Web Snippets

Valentin Jijkoun and Maarten de Rijke

ISLA, University of Amsterdam
{jijkoun,mdr}@science.uva.nl

Abstract. We describe our participation in the WebCLEF 2007 task, targeted at snippet retrieval from web data. Our system ranks snippets based on a simple similarity-based centrality, inspired by the web page ranking algorithms. We experimented with retrieval units (sentences and paragraphs) and with the similarity functions used for centrality computations (word overlap and cosine similarity). We found that using paragraphs with the cosine similarity function shows the best performance with precision around 20% and recall around 25% according to human assessments of the first 7,000 bytes of responses for individual topics.

1 Introduction

The WebCLEF 2007 task¹ differed substantially from the previous editions (2005–2006 of WebCLEF). Rather than retrieving a ranked list of web documents relevant to a topic, in the 2007 setup, systems were asked to return a ranked list of *snippets* (character spans) extracted from the top 1,000 web documents identified using the Google web search engine. The definition of the retrieval unit (snippet) was left up to a system, and thus the task is targeting *information retrieval* rather than *document retrieval*.

The remainder of this paper is organized as follows. We describe the WebCLEF 2007 task and topics in Section 2, present the architecture of our system in Section 3, describe our three runs, evaluation measures and evaluation results in Section 4, and conclude in Section 5.

2 Task and Topics

In WebCLEF 2007 for each topic systems are provided with the following information:

- topic title (e.g., *Big Bang Theory*);
- description of the information need (e.g., *I have to make a presentation about Big Bang Theory for undergraduate students. I assume that the students have some basic knowledge of physics.*);
- languages in which information can be returned;

¹ URL: <http://ilps.science.uva.nl/WebCLEF/WebCLEF2007>

- known sources: URLs and content of pages already “known” to the topic author;
- a list of web pages (original and text format) retrieved using Google with queries provided by the topic author (e.g., *Big Bang*); for each query, at most 1,000 pages are included in the list.

The task of a system is to return a ranked list of text spans from the provided web pages that, together, would satisfy the user’s information need.

Task organizers provided two development topics and 30 test topics.

3 System Architecture

For each topic, our system used only text versions of the web documents. On the one hand, the decision not to use the original versions of the documents (HTML, PDF, Postscript, etc.) led to some noise in the output of the system. In the text versions, the text encoding was often broken, which was especially problematic for non-English documents (the task included Spanish and Dutch topics and pages). Moreover, in cases where an original document was a double-column PDF, in the corresponding text version, the lines of the columns were often intervened, making the text version hardly readable for humans. For some of the original documents (e.g., for Word files) text versions were missing, and therefore our system did not use these documents at all. On the other hand, using only text version simplified the data processing considerably:

- no sophisticated content extraction had to be developed, and
- the text versions often preserved some text layout of the original pages (e.g., paragraph starts), which we used to detect suitable snippet boundaries.

Given a topic, our system first identifies *candidate snippets* in the source documents by simply splitting the text of the documents into sentences (using punctuation marks as separators) or into paragraphs (using empty lines as separators). The same snippet extraction method is applied to the text of the “known” pages for the topic, resulting in a list of *known snippets*. We ignored candidate and known snippets shorter than 30 bytes.

3.1 Ranking Snippets

We rank the candidate snippets based on *similarity-based centrality*, which is a simplified version of the graph-based snippet ranking of [2], inspired by the methods for computing authority of the Web pages, such as PageRank and HITS [4]. For each candidate snippet we compute a *centrality score* by summing similarities of the snippet with all other candidate snippets. Then, to avoid assigning high scores to snippets containing information that is already known to the user, we subtract from the resulting centrality score similarities of the candidate snippet with all known snippets. As a final step, we remove from consideration candidate snippets whose similarity to one of the known snippets is higher than a threshold. The pseudocode for this calculation is shown below:

```

let  $c_1 \dots c_n$  be candidate snippets
let  $k_1 \dots k_m$  be known snippets
for each candidate snippet  $c$ 
  let  $score(c) = 0$ 
  for each candidate snippet  $c'$ 
    let  $score(c) = score(c) + sim(c, c')$ 
  for each known snippet  $k$ 
    let  $score(c) = score(c) - sim(c, k)$ 
  for each known snippet  $k$ 
    if  $sim(c, k) > sim_{max}$ 
      let  $score(c) = 0$ 

```

Finally, the candidate snippets are ranked according to $score(\cdot)$ and top snippets are returned so that the total size of the response is not larger than 10,000 bytes.

3.2 Similarity between Snippets

A key component of our snippet ranking method is the snippet similarity function $sim(x, y)$. Similarly to [3], we conducted experiments with two versions of the similarity function: one based on word overlap and one based on the cosine similarity in the vector space retrieval model. Specifically, for two text snippets, *word overlap* similarity is defined using the standard Jaccard coefficient on snippets considered as sets of terms:

$$sim_{wo}(x, y) = \frac{|x' \cap y'|}{|x' \cup y'|},$$

where x' and y' are sets of non-stopwords of snippets x and y respectively.

The *vector space* similarity between two snippets is defined as the cosine of the angle between the vector representations of the snippets computed using the standard TF.IDF weighting scheme:

$$sim_{vs}(x, y) = \frac{\vec{x} \cdot \vec{y}}{\sqrt{\vec{x} \cdot \vec{x}} \sqrt{\vec{y} \cdot \vec{y}}}.$$

Here, $\vec{a} \cdot \vec{b}$ denotes the scalar product of vectors $\vec{a}$ and $\vec{b}$.

Components of the vectors correspond to distinct non-stopword terms occurring in the set of candidate snippets. For a term t , the value of the component $\vec{a}$ is defined according to the TF.IDF weighting scheme:

$$\vec{a}(t) = TF(a, t) \cdot \log \left(\frac{n}{|\{c_i : t \in c_i\}|} \right).$$

Here, $TF(a, t)$ is the frequency of term t in snippet a and $c_1, \dots, c_n$ are all candidate snippets.

Both versions of the similarity function produce values between 0 and 1. The similarity threshold sim_{max} for detecting near-duplicates is selected based on manual assessment of duplicates among candidate snippets for the development topics.

4 Submitted Runs and Evaluation Results

Our goal was to experiment with the units of snippet retrieval and with similarity functions. We submitted three runs:

- **UvA sent wo** – snippets defined as sentences, word overlap used for ranking;
- **UvA par wo** – snippets defined as paragraphs, word overlap for ranking;
- **UvA par vs** – paragraphs, vector space similarity.

The evaluation measures used for the task were character-based precision and recall, based on human assessments of the first 7,000 bytes of system’s response. *Precision* is defined as the length of the character spans in the response identified by humans as relevant, divided by the total of the response (limited to 7,000 characters). *Recall* is defined as the length of spans in the response identified as relevant, divided by the total length of all distinct spans identified as relevant for the responses submitted by all systems.

The evaluation results for the three runs are shown below:

Run	Precision	Recall
UvA sent wo	0.0893	0.1133
UvA par wo	0.1959	0.2486
UvA par vs	0.2018	0.2561

The results indicate that paragraphs provide better retrieval units and using a more sophisticated similarity function based on the vector space model has a slight positive effect on the performance. Unfortunately, we did not have enough time to analyse performance of the versions of the system per topic or to check whether the improvement with the vector space similarity function is significant.

Overall, we believe that the paragraph-based runs may serve as a reasonable baseline for the WebCLEF task: around 1/5 of the returned character content is considered relevant by human assessors. At the same time, such performance is probably not sufficient for a real-life information retrieval system.

5 Conclusions

We have described our participation in the WebCLEF 2007 snippet retrieval task. In our submission we experimented with retrieval units (sentences vs. paragraphs) and with similarity functions used for semantic centrality computations (word overlap vs. cosine similarity). We found that using paragraphs with the cosine similarity function shows the best performance with precision around 20% and recall around 25% according to human assessments of the first 7,000 bytes of per-topic responses.

Detailed analysis of the performance of the runs is part of our immediate agenda for future work. Another interesting direction for further study is the similarity model suitable for short snippets. The vector space model that we use in this paper is not necessarily the best option. However, it has been shown (see, e.g., [1,2]) that more sophisticated models do not necessarily lead to improvements when working with short text fragments.

Acknowledgements

The research presented in this paper 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. We are grateful to all participants and assessors of the WebCLEF 2007 task.

References

1. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: SIGIR 2003, pp. 314–321 (2003)
2. Adafre, S.F., Jijkouni, V., de Rijke, M.: Fact discovery in Wikipedia. In: IEEE/WIC/ACM International Conference on Web Intelligence 2007 (2007)
3. Jijkoun, V., de Rijke, M.: Recognizing textual entailment: Is lexical similarity enough? In: Dagan, I., Dalche, F., Quinonero Candela, J., Magnini, B. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 449–460. Springer, Heidelberg (2006)
4. Liu, B.: Web Data Mining. Exploring Hyperlinks, Contents and Usage Data. Springer, Heidelberg (2006)

Using Web-Content for Retrieving Snippets

Okky Hendriansyah, Tri Firgantoro, and Mirna Adriani

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
okky@ui.edu, mirna@cs.ui.ac.id

Abstract. We report on our participation in the web task of the 2007 Cross-Language Evaluation Forum (CLEF). We compared the results of snippet extraction based on topic title, ordered window, and unordered window. The precision and recall of the snippet extraction based on topic title was the best compared to those of the ordered window and unordered window.

1 Introduction

The availability of a huge amount of web documents on the Internet has driven research in identifying useful characteristics for extracting important information from web documents. Web retrieval is one of the research topics that concentrate in retrieving web documents [1, 2]. Web documents contain special characteristics that regular documents do not have, such as links which refer to other documents or by referred to by other documents; anchor text, URL representing storage structure, etc. [3]. Using these web page characteristics to retrieve web documents have been demonstrated to produce good results [3].

Taking advantage of the web document characteristics is just one of the stages in web retrieval. Users often need to generate snippets once the relevant web documents have been found [4]. A snippet contains important information that related to the query produced by the user. A snippet is extracted from the web documents in various lengths. The focus of the web task of the Cross-Language Evaluation Forum 2007 is to identify the characteristics of the web pages that are important to producing the snippets.

2 The Snippet Scoring

The web document entries in the result list are usually accompanied by their snippets that are considered useful for the user's information need. The snippets are scored by its similarity with the topic title. This can be computed using one of three techniques:

1. The first technique calculates the similarity between topic title and snippets. The snippets that contain more words from the topic titles are considered more important than other snippets. If a word in the snippets also appears in the topic title, then the weight of the word is multiplied by two. The score of a snippet is the sum of weights of words that appeared on the snippet.

2. The second technique calculates the score of the snippets by calculating the number of words appeared on a certain window. The window's length is ten words. Words that are in the topic title must appear in the window in similar order as the topic title. Then the score of the snippet is the total weight of all the words in the window.

3. The third technique calculates the score of the snippets by calculating the number of words appeared in a certain window. The window's length is ten words. The words on the topic title can appear on the window in any order. Then the score of the snippet is the total weight of all the words on the window.

The snippets are then ranked based on their scores. The snippet that has the highest score is considered as an important text for that topic title.

3 Experiment

The web document task searches for web pages in a number of languages. In these experiments, we use only the textual part of the documents. We remove all the stop-words from the web documents. We do not apply word stemming to the web documents since our previous work shows that doing so did not help in improving the retrieval effectiveness. Then the content of the documents is split into passages that contain two sentences each. The snippet is ranked based on the score that is calculated based on the weight of the words. The weight of each word in the web document is calculated using the *tf.idf* formula [5].

4 Results

The evaluation is performed using the precision and recall as commonly done in Information Retrieval researches:

- Precision is the number of characters that belong to at least one span linked to a nugget, divided by the total character length of the system's response.
- Recall is the sum of character lengths of all spans in R linked to nuggets, divided by the total sum of span lengths in the responses for T in all submitted runs.

The result is shown in Table 1. The best retrieval effectiveness in getting the snippets was achieved by giving more weights to words that appear in the topic title. By considering the topic title words that appear within certain word length in any order

Table 1. Evaluation result using words that appear on the title, ordered window, and unordered window

Run	Recall	Precision
Title-based	0.114	0.031
Ordered window based	0.109	0.030
Unordered window based	0.092	0.026

has decreased the precision and recall. However, by considering topic title words that appear in a window with similar order as in the topic title, better precision and recall is achieved.

5 Summary

Our results demonstrate that considering the words that appear in the topic title resulted in the best snippet score compared to using windows with ordered and unordered words. We hope to improve our results in the future by exploring still other methods.

References

1. Hawking, D.: Overview of the TREC-9 Web Track. In: The 10th Text Retrieval Conference (TREC-10), NIST Special Publication (2001)
2. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of ACM SIGIR 1998, Melbourne, Australia (August 1998)
3. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, New York (1999)
4. Jijkoun, V., de Rijke, M.: Overview of the WebCLEF. In: CLEF 2007 Working Notes (2007)
5. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)

GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview

Thomas Mandl¹, Fredric Gey², Giorgio Di Nunzio³, Nicola Ferro³, Ray Larson², Mark Sanderson⁴, Diana Santos⁵, Christa Womser-Hacker¹, and Xing Xie⁶

¹Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de, womser@uni-hildesheim.de

²University of California, Berkeley, CA, USA
gey@berkeley.edu, ray@sims.berkeley.edu

³Department of Information Engineering, University of Padua, Italy
dinunzio@dei.unipd.it, ferro@dei.unipd.it

⁴Department of Information Studies, University of Sheffield, Sheffield, UK
m.sanderson@sheffield.ac.uk

⁵Linguateca, SINTEF ICT, Norway
Diana.Santos@sintef.no

⁶Microsoft Research Asia, Beijing, China
Xingx@microsoft.com

Abstract. GeoCLEF ran as a regular track for the second time within the Cross Language Evaluation Forum (CLEF) 2007. The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR): retrieval for topics with a geographic specification. GeoCLEF 2007 consisted of two sub tasks. A search task ran for the third time and a query classification task was organized for the first. For the GeoCLEF 2007 search task, twenty-five search topics were defined by the organizing groups for searching English, German, Portuguese and Spanish document collections. All topics were translated into English, Indonesian, Portuguese, Spanish and German. Several topics in 2007 were geographically challenging. Thirteen groups submitted 108 runs. The groups used a variety of approaches. For the classification task, a query log from a search engine was provided and the groups needed to identify the queries with a geographic scope and the geographic components within the local queries.

1 Introduction

GeoCLEF¹ is the first track in an evaluation campaign dedicated to evaluating geographic information retrieval systems. 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 newspaper collections. GeoCLEF 2005 was run as a pilot track and in 2006, GeoCLEF was a regular CLEF track. GeoCLEF has continued

¹ <http://www.uni-hildesheim.de/geoclef/>

to evaluate retrieval of documents with an emphasis on geographic information retrieval from text. Geographic search requires the combination of spatial and content based relevance into one result. Many research and evaluation issues surrounding geographic mono- and bilingual search have been addressed in GeoCLEF.

GeoCLEF was a collaborative effort by research groups at the University of California, Berkeley (USA), the University of Sheffield (UK), the University of Hildesheim (Germany) and Linguatca (Norway and Portugal). Thirteen research groups (17 in 2006) from a variety of backgrounds and nationalities submitted 108 runs (149 in 2006) to GeoCLEF.

For 2007, Portuguese, German and English were available as document and topic languages. There were two Geographic Information Retrieval tasks: monolingual (English, German and Portuguese) where both topics and queries were in a single language and bilingual (topics in language X to documents in language Y, where X or Y was one of English, German or Portuguese, and X could in addition be Spanish or Indonesian).

In the three editions of GeoCLEF so far, 75 topics with relevance assessments have been developed. Thus, GeoCLEF has developed a standard evaluation collection which supports long-term research.

Table 1. GeoCLEF test collection – collection and topic languages

GeoCLEF Year	Collection Languages	Topic Languages
2005 (pilot)	English, German	English, German
2006	English, German, Portuguese, Spanish	English, German, Portuguese, Spanish, Japanese
2007	English, German, Portuguese	English, German, Portuguese, Spanish, Indonesian

Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Many documents contain some kind of spatial reference which may be important for IR. For example, to retrieve, rank and visualize search results based on a spatial dimension (e.g. “find me news stories about bush fires near Sidney”).

Many challenges of geographic IR involve geographical references (geo-references). Documents contain geo-references expressed in multiple languages which may or may not be the same as the query language. For example, the city *Cape Town* (English) is also *Kapstadt* (German), *Cidade do Cabo* in Portuguese and *Ciudad del Cabo* (Spanish). Queries with names may require an additional translation step to enable successful retrieval. Depending on the language and the culture, translation may not be helpful in some cases. For example, the word *new* within *New York* is often translated in Spanish (*Nueva York*) and Portuguese (*Nova Iorque*), but never in German. On some occasions, names may be changed and a recent modification may not be well reflected within a foreign collection. E.g. there were still references to the German city *Karl-Marx-Stadt* in Spain after it had been renamed to *Chemnitz* in 1990. Geographical references are often ambiguous (e.g. there is a *St. Petersburg* also in Florida and Pennsylvania in the United States).

The query parsing (and classification) task was offered for the first time at GeoCLEF 2007. This task was dedicated to identifying geographic queries within a log

file from the MSN search engine. A log of real queries was provided. Some were labeled as training data and some as test data. The task required participants to find geographic queries within the set and to further mark the geographic entities within the query. The task is briefly described in section 5.

2 GeoCLEF 2007 Search Task

Search is the main task of GeoCLEF. The following sections describe the test design adopted by GeoCLEF.

2.1 Document Collections Used in GeoCLEF 2007

The document collections for this year's GeoCLEF experiments consists of newspaper and newswire stories from the years 1994 and 1995 used in previous CLEF ad-hoc evaluations [1]. The Portuguese, 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 newswire agency *Schweizer Depeschen Agentur* (SDA, 1994/95). For Portuguese, GeoCLEF 2007 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 with documents. The Portuguese collections are also distributed for IR and NLP research by Linguatca as the CHAVE² collection [2].

Table 2. GeoCLEF 2007 test collection size

Language	English	German	Portuguese
Number of documents	169,477	294,809	210,734

In all 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 and contained no semantic location-specific information.

2.2 Generating Search Topics

A total of 25 topics were generated for this year's GeoCLEF (GC51 - GC75). Topic creation was shared among the three organizing groups, who all utilized the DIRECT

² <http://www.linguatca.pt/CHAVE/>

System provided by the University of Padua [3]. A search utility for the collections was provided within DIRECT to facilitate the interactive exploration of potential topics. Each group created initial versions of nine proposed topics in their language, with subsequent translation into English. Topics are meant to express a natural information need which a user of the collection might have [4]. These candidates were subsequently checked for relevant documents in the other collections. In many cases, topics needed to be refined. For example, the topic candidate *honorary doctorate degrees at Scottish universities* was expanded to topic GC53 *scientific research at Scottish universities* due to an initial lack of relevant documents in the German and Portuguese collections. Relevant documents were marked within the DIRECT system. After intensive discussion, a decision was made about the final set of 25 topics. Finally, all missing topics were translated into Portuguese and German and all translations were checked. The following section will discuss the creation of topics with spatial parameters for the track.

The organizers continued the efforts of GeoCLEF 2006 aimed at creating a geographically challenging topic set. This means that explicit geographic knowledge should be necessary in order for the participants to successfully retrieve relevant documents. Keyword-based approaches should not be favored by the topics. While many geographic searches may be well served by keyword approaches, others require a profound geographic reasoning. We speculate that for a realistic topic set where these difficulties might be less common, most systems could perform better.

In order to achieve that, several difficulties were explicitly included into the topics of GeoCLEF 2006 and 2007:

- ambiguity (*St. Paul's Cathedral*, exists in London and São Paulo)
- vague geographic regions (*Near East*)
- geographical relations beyond IN (*near Russian cities, along Mediterranean Coast*)
- cross-lingual issues (*Greater Lisbon* , Portuguese: *Grande Lisboa* , German: *Großraum Lissabon*)
- granularity below the country level (*French speaking part of Switzerland, Northern Italy*)
- complex region shapes (*along the rivers Danube and Rhine*)

However, it was difficult to develop topics which fulfilled all criteria. For example, local events which allow queries on a level of granularity below the country often do not lead to newspaper articles outside the national press. This makes the development of cross-lingual topics difficult.

For English topic generation, topics were initially generated by Mark Sanderson and tested on the DIRECT system. Additional consultation was conducted with other members of the GeoCLEF team to determine if the topics had at least some relevant documents in the German and Portuguese collections. Those found to have few such documents were altered in order ensure that at least some relevant documents existed for each topic.

The German group at Hildesheim started with brain storming on interesting geographical notions. Challenging geographic notions below the country granularity were procured. We came up with German speaking part of Switzerland, which is a vaguely defined region. A check in the collection showed that there were sport events,

but not enough to specify a sport discipline. Another challenge was introduced with Nagorno-Karabakh which has many spelling variants.

The Portuguese topics were chosen in a way similar to the one suggested for the choice of ad-hoc topics in previous years [2]. The tripartite division among international, European and national, however, was reduced to national vs. international because we did not consider European as a relevant category (given that neither Portuguese nor English language newspaper collections used in CLEF are totally based in Europe): so, we chose some culturally-bound topics (Senna, crime in Grande Lisboa), some purely international or global (sharks and floods) and some related to specific regions (because of the geographic relevance to GeoCLEF).

In all cases, but especially for those focusing on a particular region (inside or outside the national borders covered by any newspaper collection), we tried to come up with a sensible user model: either a prospective tourist (St. Paul's or Northern Italy) or a cub reporter (Myanmar human rights violation or casualties in the Himalaya). In some cases, we managed to create topics whose general relevance could be either, although naturally the choices would be different for the different kind of users – consider the case of navigation in the Portuguese islands, both relevant for a tourist and for a journalist discussing the subject.

We were also intent on trying some specifically known geographically ambiguous topics, such as St. Paul's or topics where the geographical names were ambiguous with non geographic concepts, such as Madeira (means wood in Portuguese and can also mean a kind of wine).

All the topics were then tried out in the CHAVE collection, encoded in CQP [5] and available for Web search through the AC/DC³ project [6] at in order to estimate the number of possible hits. In general, there were very few hits for all topics, as can be appreciated by the number of relevant documents per topic found in the Portuguese pool (see Table 5).

The translation of the topics leads to new challenges. One of the English topics about the Scottish town, St. Andrews, was judged to be challenging as it was more ambiguous than in English, because Santo André also denotes a village in Portugal and a city in Brazil. So this is a case where depending on the language the kind of results expected is different. While we are not defending a user model where this particular case would be relevant, we are showing that a mere topic translation (as might be effected by a cross lingual system) would not be enough if one were interested in the Scottish St. Andrews alone.

Another interesting remark is the use of the word “continent”, which is very much context dependent and again therefore cannot be translated simply from “continent” to “continente”, because depending on your spatial basis the continent is different. Again this requires some clever processing and/or processing for the translation.

Finally, it appears that *perto de X* (near X, or close to X) carries in Portuguese the presupposition that X is not included, and this made us consider that we would have translated better “airports near to London” by “que servem Londres” (i.e., that are used to reach London). (Although we also used the phrase aeroportos londrinos which may also include airports inside London). On the other hand, airplane crashes close to Russian cities seemed more naturally translated by “na proximidade” and not included. We used *perto* for both, but this might have been a translation weakness.

³ <http://www.linguatca.pt/ACDC/>

2.3
Format of Topic Description

The format of GeoCLEF 2007 was the same of the one of 2006 [7], in that no markup of geographic entities in the topics was provided as had been the case in 2005 [8]. Systems were expected to extract the necessary geographic information from the topic. Two examples of full topics are shown in Figure 1.

<num>10.2452/58-GC</num> <title>Travel problems at major airports near to London</title> <desc>To be relevant, documents must describe travel problems at one of the major airports close to London.</desc> <narr>Major airports to be listed include Heathrow, Gatwick, Luton, Stanstead and London City airport.</narr> </top>	<num>10.2452/75-GC</num> <title>Violation of human rights in Burma</title> <desc>Documents are relevant if they mention actual violation of human rights in Myanmar, previously named Burma.</desc> <narr>This includes all reported violations of human rights in Burma, no matter when (not only by the present government). Declarations (accusations or denials) about the matter only, are not relevant.</narr> </top>

Fig. 1. Topics GC058 and GC075

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. In some topics, lists of relevant regions or places were given.

2.4
Several Kinds of Geographical Topics

A tentative classification for geographical topics was suggested at GIR 2006 [9] and applied at GeoCLEF2006 [7]:

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 a 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?*)

This year we kept topics of both kinds 1 and 2 as last year. The major innovation and diversity introduced in GeoCLEF 2007 were more complicated geographic restriction than at previous GeoCLEF editions. The following three difficulties were introduced:

1. by specifying complex (multiply defined) geographic relations: East Coast of Scotland; Europe excluding the Alps, main roads north of Perth, Mediterranean coast, Portuguese islands, and “the region between the UK and the Continent”;
2. by insisting on as politically defined regions, both smaller than countries, such as French speaking part of Switzerland, the Bosphorus, Northern Italy, Grande Lisboa, or larger than countries: East European countries, Africa and north western Europe;
3. by having finer geographic subjects, such as lakes, airports, F1 circuits, and even one cathedral as place.

2.5 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 natural language processing (NLP) processing to extract place and topological clues from the texts and queries. Specific techniques used included:

- Ad-hoc techniques (weighting, probabilistic retrieval, language model, blind relevance feedback)
- Semantic analysis (annotation and inference)
- Geographic knowledge bases (Gazetteers, thesauri, ontologies)
- Text mining
- Query expansion techniques (e.g. geographic feedback)
- Geographic Named Entity Extraction (LingPipe, GATE, etc.)
- Geographic disambiguation
- Geographic scope and relevance models
- Geographic relation analysis
- Geographic entity type analysis
- Term expansion using WordNet
- Part-of-speech tagging.

2.6 Relevance Assessment

English assessment was shared by Berkeley and Sheffield Universities. German assessment was done by the University of Hildesheim and Portuguese assessment by Linguatca. The DIRECT System [3] was utilized for assessment. The system provided by the University of Padua allowed the automatic submission of runs by participating groups and supported assembling the GeoCLEF assessment pools by language.

2.6.1 English Relevance Assessment

English relevance assessment was conducted primarily by a group of ten paid volunteers from the University of Sheffield, who were paid a small sum of money for each topic assessed. The English document pool extracted from 53 monolingual and 13 bilingual (language X to) English runs consisted of 15,637 documents to be reviewed and judged by our 13 assessors or about 1,200 documents per assessor.

Table 3. GeoCLEF English 2007 Pool

Pool Size	15,637 documents • 14,987 not relevant • 650 relevant 25 topics • about 625 documents per topic
Pooled Experiments	27 out of 66 submitted experiments • monolingual: 21 out of 53 submitted experiments • bilingual: 6 out of 13 submitted experiments
Assessors	13 assessors • about 1,200 documents per assessor

The box plot of figure 2 shows the distribution of different types of documents across the topics of the English pool. In particular, the upper box shows the distribution of the number of pooled documents across the topics; as it can be noted, the distribution is a little bit asymmetric towards topics with a higher number of pooled documents and does not present outliers. The middle box shows the distribution of the number of not relevant documents across the topics; as it can be noted, the distribution is a little bit asymmetric towards topics with a lower number of not relevant documents and does not present outliers. Finally, the lower box shows the distribution of the number of relevant documents across the topics; as it can be noted, the distribution is almost symmetric; with a median number of relevant documents around 20 per topic, but it present some outliers, which are topics with a large number of relevant documents.

2.6.2 German Relevance Assessment

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. One topic required systems to find documents which report shark attacks. Documents telling the reader that a certain area is “full of sharks” were not judged as relevant.

For other topics, implicit information in the document was used for the decision. For example, the topic sport events in German speaking Switzerland led to documents where the place of a soccer game was not mentioned, but the result was included in a standardized form which indicates that the game was played in the first city

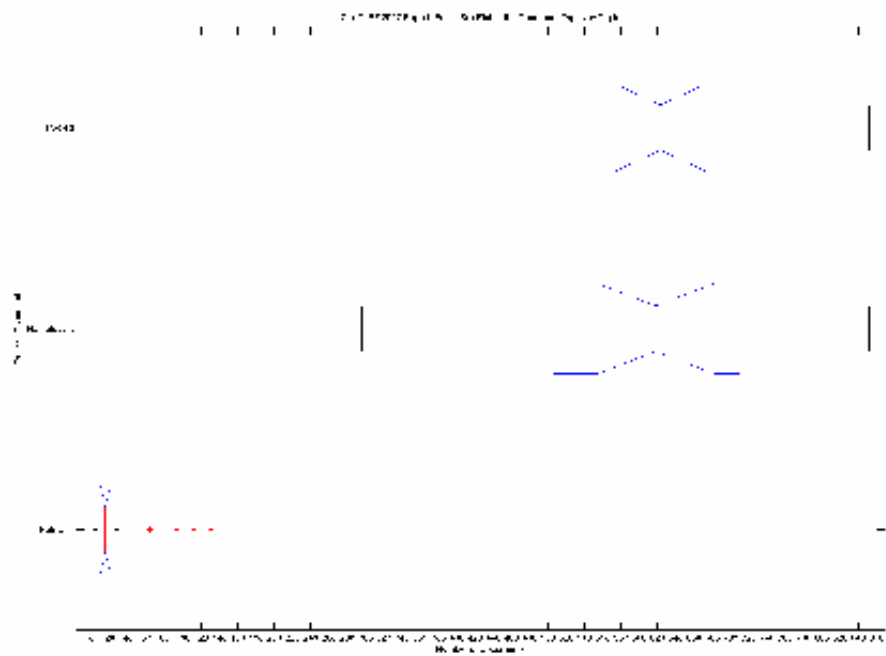


Fig. 2. GeoCLEF English 2007 Pool: distribution of the different document types

Table 4. GeoCLEF German 2007 Pool

Pool Size	15,488 documents • 14,584 not relevant • 904 relevant 25 topics • about 620 documents per topic
Pooled Experiments	24 out of 24 submitted experiments • monolingual: 16 experiments • bilingual: 8 experiments
Assessors	8 assessors • about 1,900 documents per assessor

mentioned (e.g. Lausanne - Genf 0:2, has most usually been played in Lausanne). It was also assumed that documents which report that hikers are missing in the Himalayas are relevant for the topic casualties in the Himalayas.

Many documents are at first identified as borderline cases and need to be discussed further. One topic requested topics on travel delays at London airports. One document mentioned that air travel had been delayed and some flight had to be directed to Gatwick. Because a delay at Gatwick is not explicitly mentioned, the document was regarded as not relevant.

The box plot of figure 3 shows the distribution of different types of documents across the topics of the German pool. It shows for the three sets of pooled, relevant and non relevant documents how they are distributed over the topics. This graph shows that the medium number of non relevant documents for a topic is 640. There is one topic with 300 non relevant documents which represents the minimum of the distribution. The maximum is a topic with 850 documents. The number of the topics is not given in this graph.

As it can be noted, the distribution of the pooled documents is almost symmetrical with no outliers. On the other hand, the distribution of non relevant documents is asymmetrical with a tail towards topics with a lower number of not relevant documents and does not present outliers; finally, also the distribution of the relevant documents is asymmetrical but towards topics with a greater number of relevant documents and presents outliers, which are topics with a great number of relevant documents

2.6.3 Portuguese Relevance Assessment

In addition to the problem (already reported before) that some of the news articles included in the CHAVE collection are in fact a list of “last news” which concern several different subjects (and have therefore to be read in their entirety, making it especially tiresome), we had some general problems assessing topics, which we illustrate here in detail for the “free elections in Africa” subject:

What is part of an election (or presupposed by it)? In other words, which parts are necessary or sufficient to consider that a text talks about elections: campaign, direct results, who were the winners, “tomada de posse”, speeches when receiving the power, cabinet constitution, balance after one month, after more time...

In fact, how far in time is information relevant? For example, does mention to the murder of the first democratically elected president in Ruanda qualify as text about free elections in Africa? And if elections took place and were subsequently annulated as in Argelia, do they count as elections or not? Also, how much indirectly conveyed information can be considered relevant? A text about the return of Portuguese citizens to Portugal after the (free) South African elections is about free elections in South Africa?

The decision on whether the elections were free or not might be arbitrary when this fact is not mentioned in the text. Should the juror assume anything? As in the case of a text about Uganda mentioning “voltou à Presidência no fim de 1980, pela via eleitoral” (X came back to presidency through the electoral path). Are either our knowledge or our opinions going to play a role on the relevance assessment, or we are supposed to just look at the document and not bring our own bias?

Finally, how much difference of opinions is relevant to a topic? Consider the following piece of news “Savimbi considera ilegais as eleições consideradas livres e justas pela ONU...” (Savimbi considers illegal the elections considered free and just by UN). Are we to stand with UN or with Savimbi, as far as the elections in Angola are concerned? (In our opinion, this text is very relevant to the subject, anyway, since it mentions, and discusses, precisely the issue of “free elections in an African country”).

Due to this (acknowledged) difficulty of assessing relevance for some topics, it would have been beneficial to have a pool of judges assessing the same documents and produce a relevance cline. Although this is currently not possible with the DIRECT system, it might make sense in the future, especially for more evaluative topics that involve complex issues.

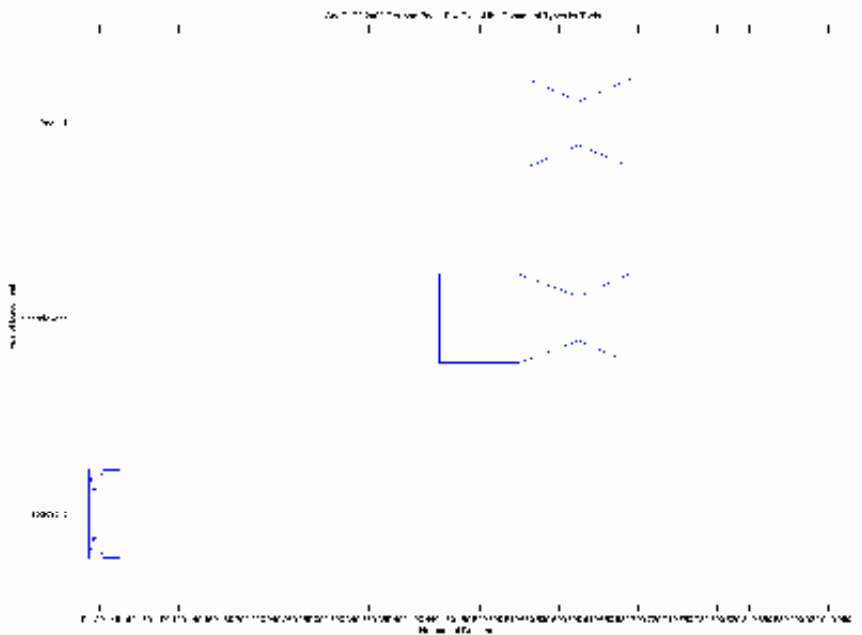


Fig. 3. GeoCLEF German 2007 Pool: distribution of the different document types

Table 5. GeoCLEF Portuguese 2007 Pool

Pool Size	15,572 documents • 14,810 not relevant • 762 relevant 25 topics • about 623 documents per topic
Pooled Experiments	18 out of 18 submitted experiments • monolingual: 11 experiments • bilingual: 7 experiments
Assessors	6 assessors • about 2,600 documents per assessor

The box plot of figure 4 shows the distribution of different types of documents across the topics of the Portuguese pool. As it can be noted the distribution of the pooled documents is a little bit asymmetrical towards topics with a lower number of pooled document and presents both upper and lower outliers, i.e. topics with many or few pooled documents; on the other hand, the distribution of not relevant documents is almost symmetrical with an outlier, which is a topic with few not relevant documents; finally, also the distribution of the relevant documents is asymmetrical towards topics with a greater number of relevant documents and presents outliers, which are topics with a great number of relevant documents.

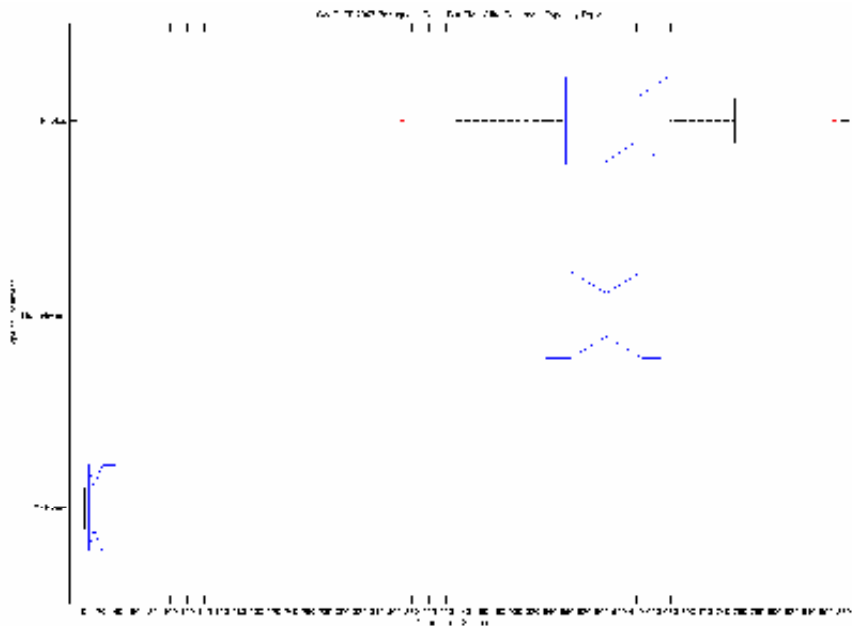


Fig. 4. GeoCLEF Portuguese 2007 Pool: distribution of the different document types

3 Results of the GeoCLEF 2007 Search Task

The results of the participating groups are reported in the following sections.

3.1 Participants and Experiments

As shown in Table 6, a total of 13 groups from 9 different countries submitted results for one or more of the GeoCLEF tasks. A total of 108 experiments were submitted.

Table 6. GeoCLEF 2007 participants – new groups are indicated by *

Participant	Institution	Country
catalunya	U.Politecnica Catalunya	Spain
cheshire	U.C.Berkeley	United States
csusm	Cal State U.- San Marcos	United States
depok*	U. Indonesia	Indonesia
groningen	U. Groningen	The Netherlands
hagen	U. Hagen-Comp.Sci	Germany
hildesheim	U. Hildesheim	Germany
icl	Imperial College London - Computing	United Kingdom
linguit*	Linguit Ltd	United Kingdom
moscow*	Moscow State U.	Russia
msasia	Microsoft Asia	China
valencia	U.Politecnica Valencia	Spain
xldb	U.Lisbon	Portugal

Five different topic languages were used for GeoCLEF bilingual experiments: German, English, Indonesian, Portuguese, and Spanish. Differently from usual, the most popular language for queries was Spanish (11 experiments out of 28 bilingual experiments); English (7 experiments) and Indonesian (6 experiments) almost tied for the second place; German (2 experiments) and Portuguese (2 experiments) tied for the third place. The number of bilingual runs by topic language is shown in Table 9.

Table 7 reports the number of participants by their country of origin.

Table 7. GeoCLEF 2007 participants by country

Country	# Participants
China	1
Germany	2
Indonesia	1
Portugal	1
Russia	1
Spain	2
The Netherlands	1
United Kingdom	2
United States	2
TOTAL	13

Table 8 provides a breakdown of the experiments submitted by each participant for each of the offered tasks.

Table 8. GeoCLEF 2007 experiments by task

Participant	Monolingual Tasks			Bilingual Tasks			TOTAL
	DE	EN	PT	X2DE	X2EN	X2PT	
catalunya		5					5
cheshire	1	1	1	3	3	3	12
csusm	6	6	5		4	4	25
depok*					6		6
groningen		5					5
hagen	5			5			10
hildesheim	4	4					8
icl		4					4
linguit*		4					4
moscow*		2					2
msasia		5					5
valencia		12					12
xldb		5	5				10
TOTAL	16	53	11	8	13	7	108

3.2 Monolingual Experiments

Monolingual retrieval was offered for the following target collections: English, German, and Portuguese. Table 10 shows the top five groups for each target collection,

Table 9. Bilingual experiments by topic language

Track	Source Language					TOTAL
	DE	EN	ES	ID	PT	
Bilingual X2DE		6	1		1	8
Bilingual X2EN	1		5	6	1	13
Bilingual X2PT	1	1	5			7
TOTAL	2	7	11	6	2	28

ordered by mean average precision. Note that only the best run is selected 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 experiment Digital Object Identifier (DOI); the mean average precision achieved by the experiment; and the performance difference between the first and the last participant.

Due to an error, the XLDB group submitted the wrong run files for monolingual Portuguese. Because of the low number of participants, this run appears among the top runs. This explains the large difference between the second and the third run in Table 10.

Figures 5 to 7 show the interpolated recall vs. average precision for the top participants of the monolingual tasks.

Table 10. Best entries for the monolingual track. Additionally, the performance difference between the best and the last (up to 5) placed group is given (in terms of mean average precision) – new groups are indicated by *.

Track	Rnk	Partner	Experiment DOI	MAP
Mono-lingual English	1 st	catalunya	10.2415/GC-MONO-EN-CLEF2007.CATALUNYA.TALPGE0IRTD2	28.5%
	2 nd	cheshire	10.2415/GC-MONO-EN-CLEF2007.CHESHIRE.BERKMOENBASE	26.4%
	3 rd	valencia	10.2415/GC-MONO-EN-CLEF2007.VALENCIA.RFIAUPV06	26.4%
	4 th	groningen	10.2415/GC-MONO-EN-CLEF2007.GRONINGEN.CLCGGEOEETD00	25.2%
	5 th	csusm	10.2415/GC-MONO-EN-CLEF2007.CSUSM.GEOMOEN5	21.3%
	Δ			33.7%
Mono-lingual German	1 st	hagen	10.2415/GC-MONO-DE-CLEF2007.HAGEN.FUHTDN5DE	25.8%
	2 nd	csusm	10.2415/GC-MONO-DE-CLEF2007.CSUSM.GEOMODE4	21.4%
	3 rd	hildesheim	10.2415/GC-MONO-DE-CLEF2007.HILDESHEIM.HIMODENE2NA	20.7%
	4 th	cheshire	10.2415/GC-MONO-DE-CLEF2007.CHESHIRE.BERKMODEBASE	13.9%
	Δ			85.1%
Mono-lingual Portuguese	1 st	csusm	10.2415/GC-MONO-PT-CLEF2007.CSUSM.GEOMOPT3	17.8%
	2 nd	cheshire	10.2415/GC-MONO-PT-CLEF2007.CHESHIRE.BERKMOPBASE	17.4%
	3 rd	xldb	10.2415/GC-MONO-PT-CLEF2007.XLDB.XLDBPT_1	3.3%
	Δ			442 %

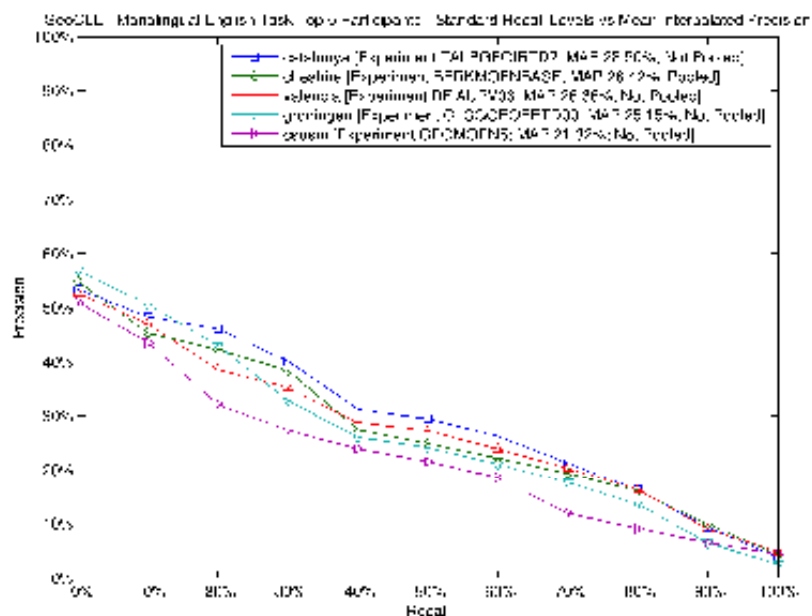


Fig. 5. Monolingual English top participants. Interpolated Recall vs. Average Precision.

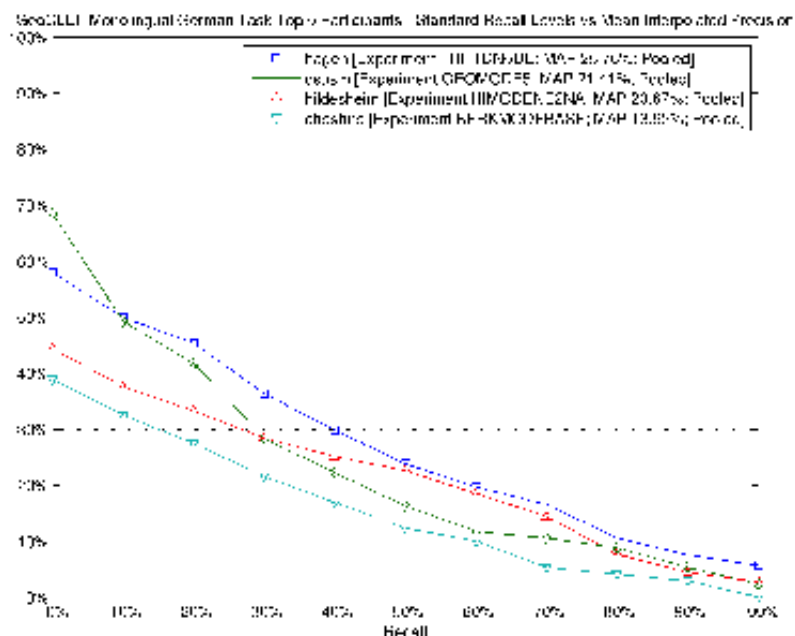


Fig. 6. Monolingual German top participants. Interpolated Recall vs. Average Precision.

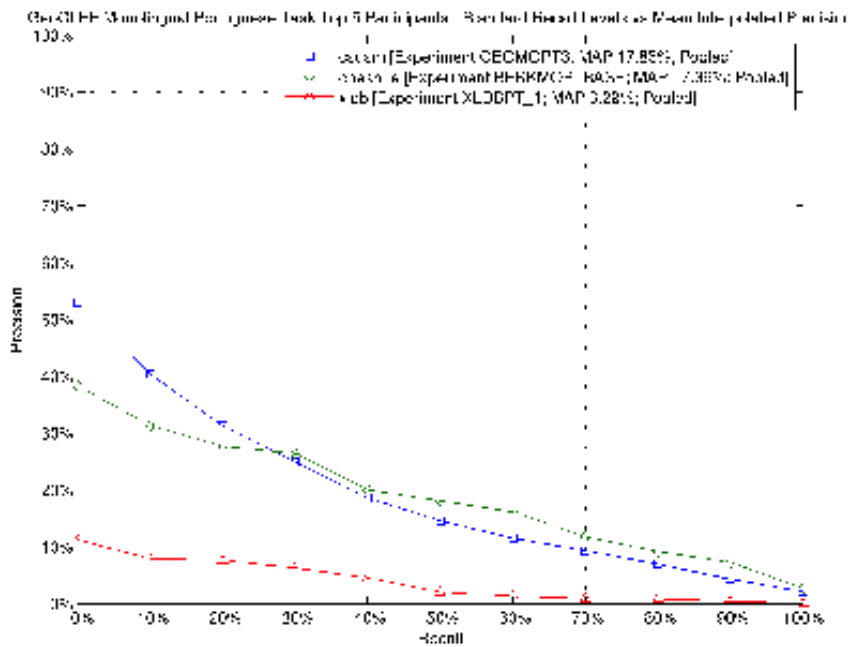


Fig. 7. Monolingual Portuguese top participants. Interpolated Recall vs. Average Precision.

3.3 Bilingual Experiments

The bilingual task was structured in three subtasks ($X \rightarrow$ DE, EN, or PT target collection). Table 11 shows the best results for this task with the same logic of Table 7. Note that the top five participants contain both “newcomer” groups and “veteran” groups.

For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines:

- $X \rightarrow$ DE: 81.1% of best monolingual German IR system
- $X \rightarrow$ EN: 77.4% of best monolingual English IR system
- $X \rightarrow$ PT: 112.9% of best monolingual Portuguese IR system

Note that there is a significant improvement for Bilingual German since CLEF 2006, when it was 70% of the best monolingual system; Bilingual English shows a small improvement, with respect to the 74% of the best monolingual system in CLEF 2006; finally, Bilingual Portuguese is quite surprising since it outperforms the monolingual and it represents a complete overturn with respect to the 47% of CLEF 2006. Figures 8 to 10 show the interpolated recall vs. average precision graph for the top participants of the different bilingual tasks.

4 Result Analysis

The test collection of GeoCLEF grew of 25 topics each year. This is usually considered the minimal test collection size to produce reliable results. Therefore, statistical

testing and further reliability analysis are performed to assess the validity of the results obtained. The range of difficulties in the topics might have led to topics more difficult and more diverse than in traditional ad-hoc evaluations. To gain some insight on this issue, a topic performance analysis was also conducted.

Table 11. Best entries for the bilingual task. The performance difference between the best and the last (up to 5) placed group is given (in terms of mean average precision) – new groups are indicated by *.

Track	Rnk.	Partner	Experiment DOI	MAP
Bilingual English	1 st	cheshire	10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDEENBASE	22.1%
	2 nd	depok*	10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGP	21.0%
	3 rd	csusm	10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN2	19.6%
	Diff.			12.5%
Bilingual German	1 st	hagen	10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDN4EN	20.9%
	2 nd	cheshire	10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIPTDEBASE	11.1%
	Diff.			88.6%
Bilingual Portuguese	1 st	cheshire	10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIENTPTBASE	20.1%
	2 nd	csusm	10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT4	5.3%
	Diff.			277.5%

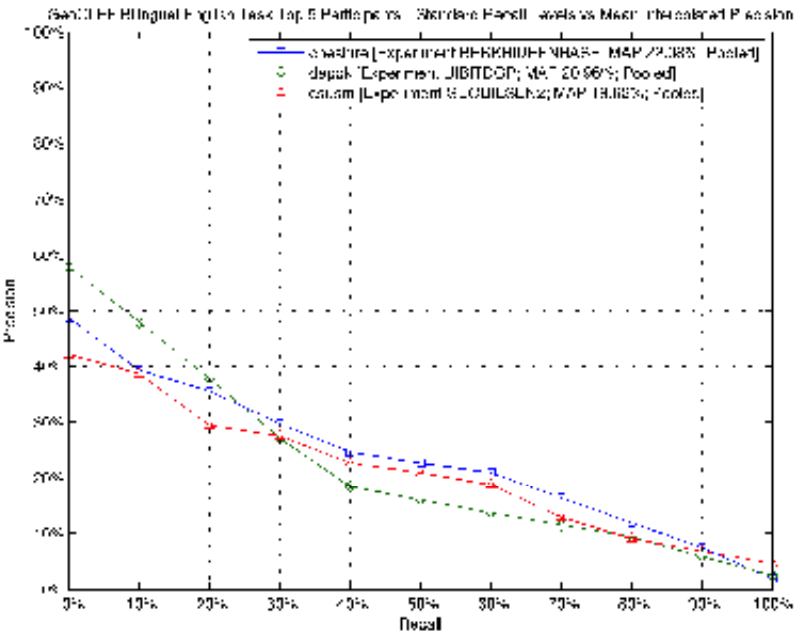


Fig. 8. Bilingual English top participants. Interpolated Recall vs Average Precision.

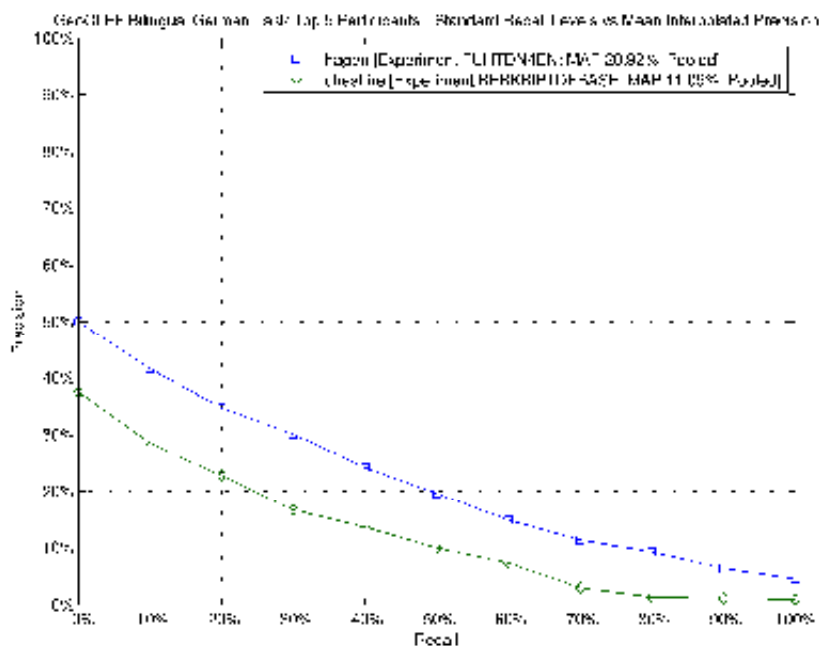


Fig. 9. Bilingual German top participants. Interpolated Recall vs Average Precision.

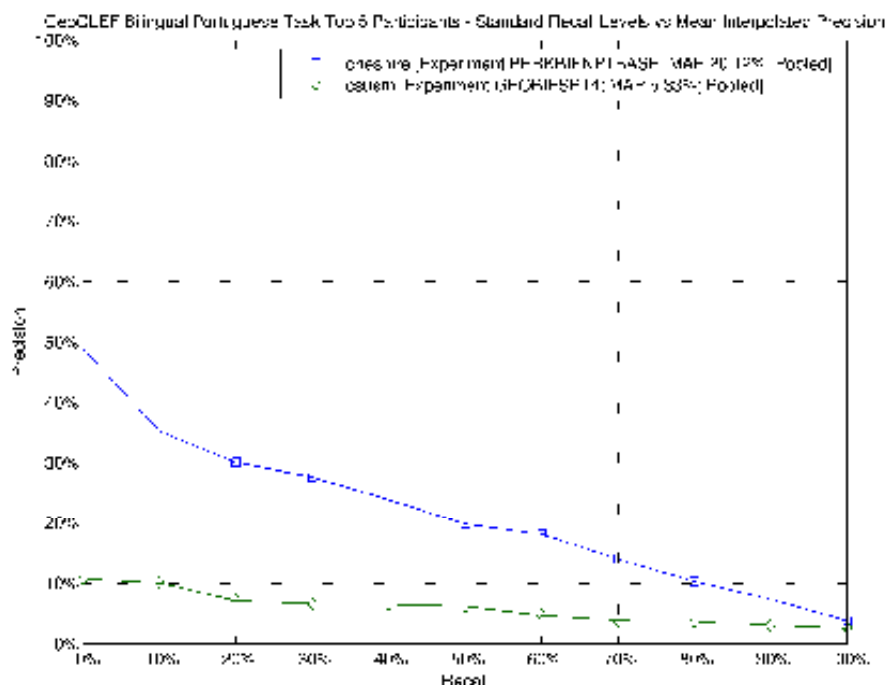


Fig. 10. Bilingual Portuguese top participants. Interpolated Recall vs Average Precision.

Table 12. 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	10	39	27	45
Monolingual German	0	13	8	14
Monolingual Portuguese	2	5	5	8
Bilingual English	1	7	10	13
Bilingual German	1	4	3	7
Bilingual Portuguese	0	2	2	3

4.1 Statistical Testing

Statistical testing for retrieval tests is intended to determine whether the order of the systems which results from the evaluation reliably measures the quality of the systems [10]. In most cases, the statistical analysis gives a conservative estimate of the upper level of significance [11]. 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.

Table 12 shows the results of the Lilliefors test before and after applying the Tague-Sutcliffe transformation. The results of the statistical analysis are shown in tables 13-18. Again, it is necessary to point out that among the few runs for monolingual Portuguese, one group were submitted with errors.

4.2 Stability Analysis

As for many other information retrieval evaluations, the variance between topics is much larger than between the systems. This fact has led doubts about the validity and reliability of tests in information retrieval. Since the variance between topics is so large, the results can depend much on the arbitrary choice of topics.

To measure this effect, a method which uses simulations with sub sets of the original topic set has been established [12]. The simulation uses smaller sets of topics and compares the resulting ranking of the systems to the ranking obtained when using all topics. If the systems are ranked very differently when only slightly smaller sets are used, the reliability is considered as small. The rankings can be compared by counting the number of position changes in the system ranking (swap rate). For GeoCLEF, such a simulation has been carried out as well. The rankings have been compared by a rank correlation coefficient. It can be observed that the system ranking remains stable even until topic sets of size 11 which is less than half of the original topic set. The correlation remains above 80% and even 90% depending on the sub task. This stability is surprisingly high and shows that the GeoCLEF results are considerably reliable.

Table 13. Monolingual German: experiment groups according to the Tukey T Test

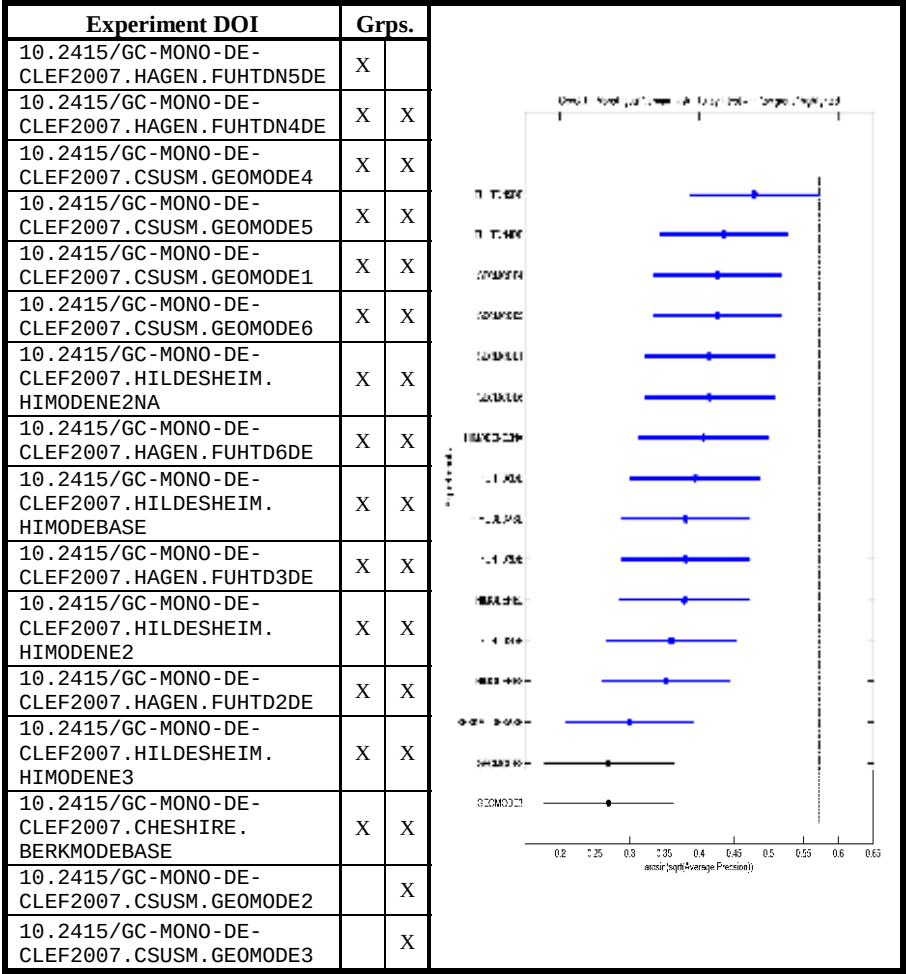


Table 14. Monolingual English: experiment groups according to the Tukey T Test. Experiment DOI is preceded by *10.2415/GC-MONO-EN-CLEF2007*.

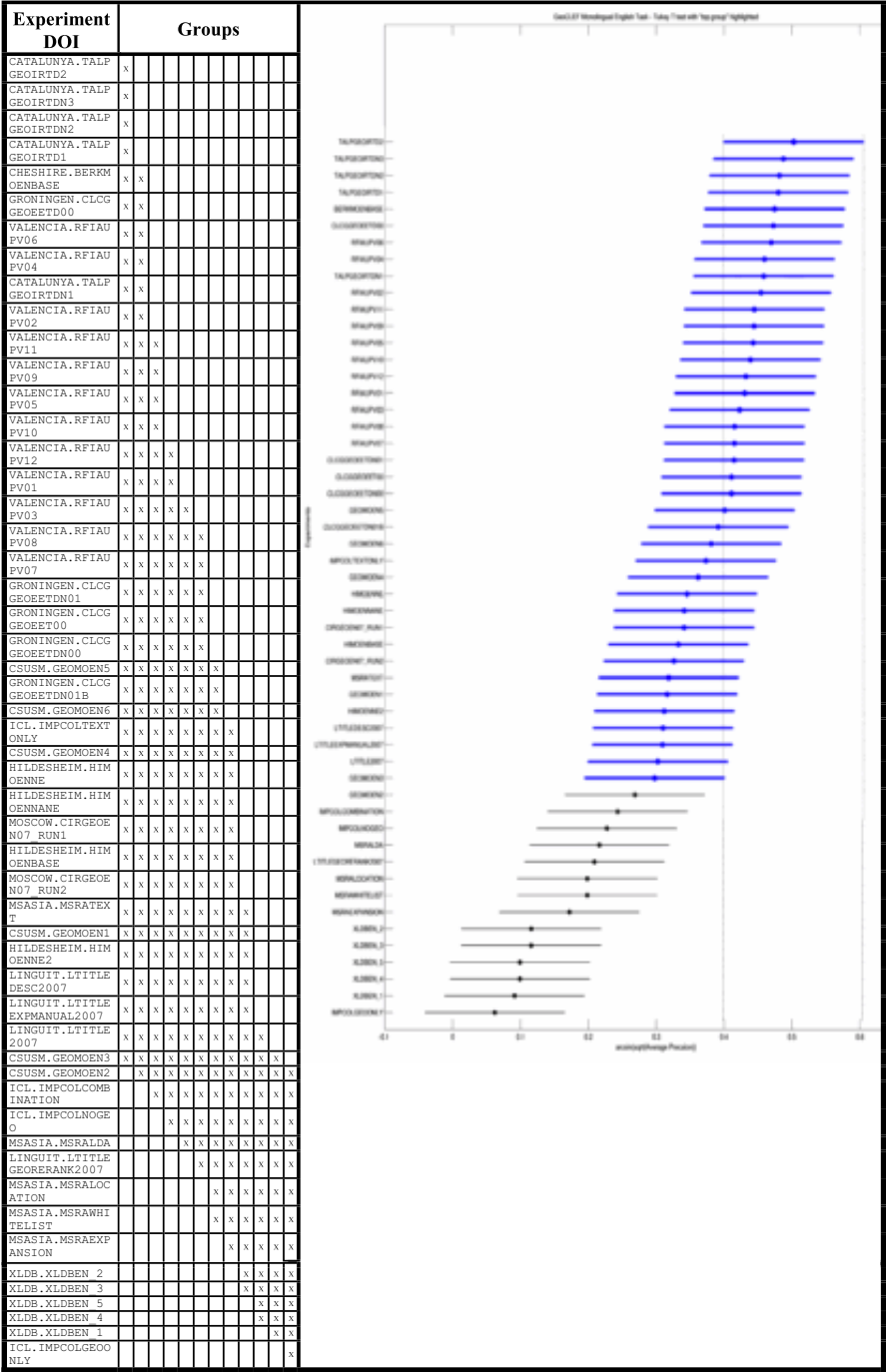


Table 16. Bilingual English: experiment groups according to the Tukey T Test

Experiment DOI	Grps
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDEENBASE	X
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIESENBASE	X
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGPGEOFB	X
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGP	X
10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIPTENBASE	X
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN2	X
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITDGPFF5	X
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN3	X
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGPPF5	X
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGP	X
10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UIBITGPGEOFB	X
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN1	X
10.2415/GC-BILI-X2EN-CLEF2007.CSUSM.GEOBIESEN4	X

ConcLRF-Rilegon English Basic Tuples T test with top-group: EnglishBasic

Experiment	Mean (Dot)	Min (Line Start)	Max (Line End)
BERKBIDEENBASE	0.28	0.22	0.38
BERKBIESENBASE	0.28	0.22	0.38
UIBITDGPGEOFB	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
BERKBIPTENBASE	0.28	0.22	0.38
GEOBIESEN2	0.28	0.22	0.38
GEOBIESEN3	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38
UIBITDGP	0.28	0.22	0.38

Table 17. Bilingual English: experiment groups according to the Tukey T Test

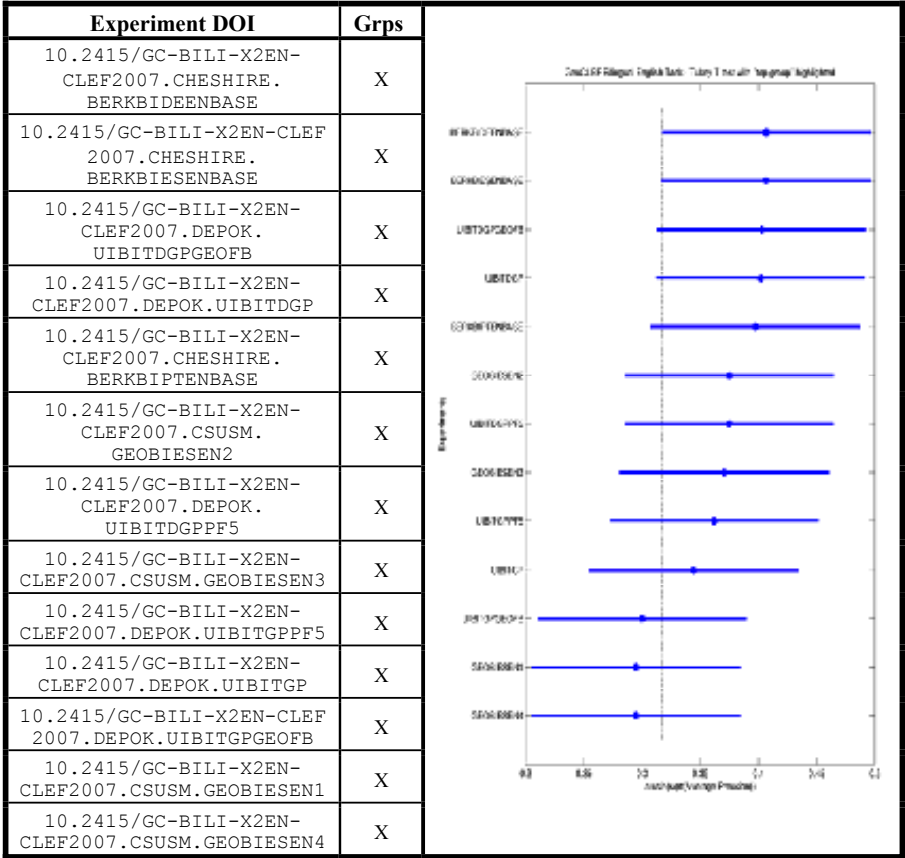


Table 18. Bilingual German: experiment groups according to the Tukey T Test

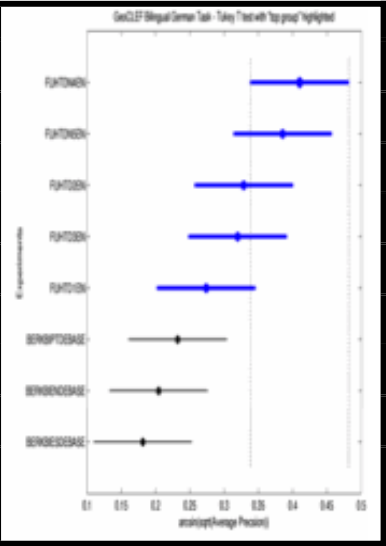
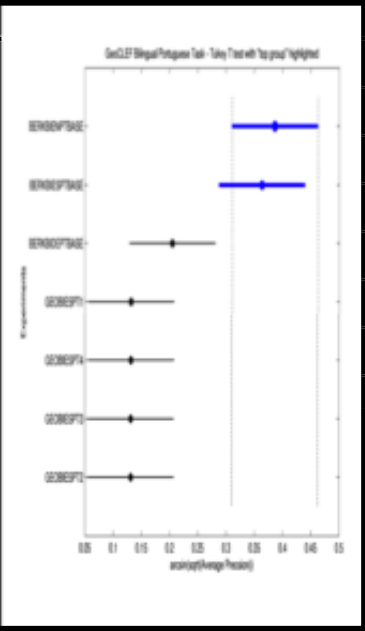
Experiment DOI	Grps			
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDN4EN	X			
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDN5EN	X			
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTD2EN	X	X		
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTD3EN	X	X	X	
10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTD1EN	X	X	X	
10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIPTDEBASE		X	X	
10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIEDEBASE		X	X	
10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIESDEBASE			X	

Table 19. Bilingual Portuguese: experiment groups according to the Tukey T Test

Experiment DOI	Grps			
10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIEPTBASE	X			
10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIESPTBASE	X			
10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIDEPTBASE			X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT1			X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT4			X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT3			X	
10.2415/GC-BILI-X2PT-CLEF2007.CSUSM.GEOBIESPT2			X	

5 Query Classification Task

The query parsing and classification task was offered for the first time at GeoCLEF 2007. It was dedicated to identifying geographic queries within a log file from the MSN search engine. This task has been organized by Xie Xing from Microsoft Research Asia. The task is of high practical relevance to GeoCLEF and the real log data is of great value for research.

The task required participants to find the geographic entity, the relation type and the non geographic topic of the query. In details, the systems needed to find the queries with a geographic scope, extract the geographic component (where), extract the type of the geographic relation (e.g. in, north of) and extract the topic of the query (what component). In addition, the systems were required to classify the query type. The classes defined were information, yellow page and map. For a query Lottery in Florida, for example, the systems were required to respond that this is a geographic query of the type information, return Florida as the where-component, lottery as the what component and extract in as the geographic relation. There were 27 geographic relations given.

For this task, a log of 800,000 real queries was provided. Out of these, 100 were labeled as training data and 500 were assessed as test data. The labeling was carried out by three Microsoft employees. They reached a consensus on each decision. In the randomly chosen and manually cleansed set, there were 36% non local queries. The geographic queries comprised 16% map queries, 29% yellow page type queries and 19% information (ad-hoc type) queries.

The results were analyzed by calculating the recall, the precision and a combined F-Score for the classification task. The task attracted six participating groups. The performance for classifying whether a query was local or not were used as a primary evaluation measure. The results are shown in Table 19.

Table 19. Results of the Query Classification Task

Team	Recall	Precision	F1-Score
Ask.com	0.625	0.258	0.365
csusm	0.201	0.197	0.199
linguit	0.112	0.038	0.057
miracle (DAEDALUS)	0.428	0.566	0.488
catalunya	0.222	0.249	0.235
xldb	0.096	0.08	0.088

The overall results are quite low. This shows that further research is necessary. Most participants used approaches which combined heuristic rules and lists and gazetteers of geographic named entities. More details on the task design, the data, participation and evaluation results are provided in an overview paper [13].

6 Conclusions and Future Work

GeoCLEF 2007 has continued to create an evaluation resource or geographic information retrieval. Spatially challenging topics have been developed and interesting experiments have been submitted. 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 the SIGIR or CIKM conferences. All participants of GeoCLEF 2007 are invited to actively contribute to the discussion of the future of GeoCLEF.

Acknowledgments

German assessment was based entirely on volunteer effort; it was carried out by Julia Jürgens, Theresa Märkl, Ben Heuwing, Frauke Zurek, Robert Stoldt, Jens Plattfaut, and Rafael Hellmann of the University of Hildesheim. The English assessment was performed by Fredric Gey, Ray Larson, Mark Sanderson and the Sheffield student team (paid for by the EU 6th Framework project Tripod). The Portuguese documents were assessed by Paulo Rocha, Luís Costa, Luís Miguel Cabral, Susana Inácio, Maria Cláudia Freitas and Diana Santos, in the scope of the Linguatca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC. The topics were thoroughly checked by Sven Hartrumpf from the University of Hagen (Germany). The query classification task was carried out by Microsoft Research Asia without any external financial support.

References

1. Braschler, M., Peters, C.: Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval* 7(1-2), 7–31
2. Santos, D., Rocha, P.: The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) *CLEF 2004*. LNCS, vol. 3491, pp. 821–832. Springer, Heidelberg (2005)
3. 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)
4. 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)
5. Evert, S.: The CQP Query Language Tutorial (CWB version 2.2.b90) University of Stuttgart (July 10, 2005), <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/>

6. Santos, D., Eckhard, B.: Providing Internet access to Portuguese corpora: the AC/DC project. In: Gavriladou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, pp. 205–210 (2000)
7. Gey, F., Larson, R., Sanderson, M., Bishoff, 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006. LNCS*, vol. 4730, pp. 852–876. Springer, Heidelberg (2007)
8. 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, pp. 908–919. Springer, Heidelberg (2006)
9. Santos, D., Chaves, M.S.: The place of place in geographical IR. In: *GIR 2006, the 3rd Workshop on Geographic Information Retrieval, SIGIR 2006, Seattle, 10 August 2006* (presentation at, 2006), <http://www.linguateca.pt/Diana/download/acetSantosChavesGIR2006.pdf>
10. Buckley, C., Voorhees, E.: *Retrieval System Evaluation*. In: *TREC: Experiment and Evaluation in Information Retrieval*, pp. 53–75. MIT Press, Cambridge (2005)
11. Sanderson, M., Zobel, J.: *Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability*. In: *28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR) 2005*, pp. 162–169 (2005)
12. Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments? In: *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1998)*, Melbourne, Australia, pp. 307–314. ACM Press, New York (1998)
13. Li, Z., Wang, C., Xing, X., Ma, W.-Y.: Query Parsing Task for GeoCLEF 2007 Report. In: Nardi, A., Peters, C. (eds.) *Working Notes of the Cross Language Evaluation Forum (CLEF) (2007)*

Inferring Location Names for Geographic Information Retrieval

Johannes Leveling and Sven Hartrumpf

Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen), 58084 Hagen, Germany
`firstname.lastname@fernuni-hagen.de`

Abstract. For the participation of GIRSA at the GeoCLEF 2007 task, two innovative features were introduced to the geographic information retrieval (GIR) system: identification and normalization of location indicators, i.e. text segments from which a geographic scope can be inferred, and the application of techniques from question answering. In an extension of a previously performed experiment, the latter approach was combined with an approach using semantic networks for geographic retrieval. When using the topic title and description, the best performance was achieved by the combination of approaches (0.196 mean average precision, MAP); adding location names from the narrative part increased MAP to 0.258. Results indicate that 1) employing normalized location indicators improves MAP significantly and increases the number of relevant documents found; 2) additional location names from the narrative increase MAP and recall, and 3) the semantic network approach has a high initial precision and even adds some relevant documents which were previously not found. For the bilingual experiments, English queries were translated into German by the Prompt machine translation web service. Performance for these experiments is generally lower. The baseline experiment (0.114 MAP) is clearly outperformed, achieving the best performance for a setup using title, description, and narrative (0.209 MAP).

1 Introduction

In geographic information retrieval (GIR) on textual information, named entity recognition and classification play an important role to identify location names. GIR is concerned with facilitating geographically-aware retrieval of information, which typically results from identifying location names in the text and classifying them into geographic and non-geographic names.

The main goal of this paper is to investigate if GIR benefits from an approach which is not solely based on identifying proper nouns corresponding to location names. To this end, the system GIRSA (Geographic Information Retrieval by Semantic Annotation)¹ was developed. GIRSA introduces the notion of location

¹ The research described is part of the IRSAW project (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web; LIS 4 – 554975(2) Hagen, BIB 48 HGfu 02-01), which is funded by the DFG (Deutsche Forschungsgemeinschaft).

Table 1. Definition of location indicator classes

Class	Definition; Example
location adjective	adjective derived from a location name; “ <i>irisch</i> ”/‘ <i>Irish</i> ’ for “ <i>Ireland</i> ”/‘ <i>Ireland</i> ’
demonym	name for inhabitants originating from a location; “ <i>Franzose</i> ”/‘ <i>Frenchman</i> ’ for “ <i>Frankreich</i> ”/‘ <i>France</i> ’
location code	code for a location, including ISO code, postal and zip code; ‘ <i>HU21</i> ’ as the FIPS region code for ‘ <i>Tolna County, Hungary</i> ’
location abbreviation	abbreviation or acronym for a location; “ <i>franz.</i> ” for “ <i>französisch</i> ”/‘ <i>French</i> ’ (mapped to “ <i>Frankreich</i> ”/‘ <i>France</i> ’)
name variant	orthographic variant, exonym, or historic name; ‘ <i>Cologne</i> ’ for “ <i>Köln</i> ”
language	language name in a text; ‘ <i>Portuguese</i> ’ for ‘ <i>Portuguese speaking countries</i> ’ (mapped to ‘ <i>Portugal, Angola, Cape Verde, East Timor, Mozambique, Brazil</i> ’)
meta-information	document language, place of publication, place of birth for the author; such attributes can be explicitly given by Dublin core elements or similar means or can be inferred from the document
unique entity	entity associated with a geographic location, including headquarters of an organization, persons, and buildings; ‘ <i>Boeing</i> ’ for ‘ <i>Seattle, Washington</i> ’; ‘ <i>Eiffel Tower</i> ’ for ‘ <i>Paris</i> ’
location name	name of a location, including full name and short form; “ <i>Republik Korea</i> ”/‘ <i>Republic of Korea</i> ’ for “ <i>Südkorea</i> ”/‘ <i>South Korea</i> ’

indicators and the application of question answering (QA) techniques to GIR. The system is evaluated on documents and topics for GeoCLEF 2007, the GIR task at CLEF 2007.

2 Location Indicators

Location indicators are text segments from which the geographic scope of a document can be inferred. Important location indicators classes are shown in Table 1.² Typically, location indicators are not part of gazetteers, e.g. the morphological and lexical knowledge for adjectives is missing completely. Distinct classes of location indicators contribute differently in assigning a geographic scope to a document; their importance depends on their usage and frequency in the corpus (e.g. adjectives are generally frequent) and the correctness of identifying them because new ambiguities may be introduced (e.g. the ISO 3166-1 code for Tuvalu (TV) is also the abbreviation for television).

For identification and normalization of location indicators, tokens are mapped to base forms and looked up in a knowledge base. The knowledge base contains pairs of a location indicator and a normalized location name. This knowledge base was created by collecting raw material from web sources and dictionaries

² German examples are double-quoted, while English examples are single-quoted.

(including Wikipedia and an official list of state names³), which was then transformed into a machine-readable form, manually extended, and checked.

Location indicators are normalized to location names on different levels of linguistic analysis in GIRSA. Normalization consists of several stages. First, Morphological variations are identified and inflectional endings are removed, reducing location indicators to their base form. In addition, multi-word names are recognized and represented as a single term (“*Roten Meer(e)s*”/ ‘*Red Sea’s*’ → “*Rote_Meer*”/ ‘*Red_Sea*’).

In the next step, location indicators are normalized, e.g. abbreviations and acronyms are expanded and then mapped to a synset representative, e.g. equivalent location names containing diacritical marks or their equivalent non-accented characters are represented by an element of the name synset (e.g., “*Québec*” → “*Quebec*”).

Finally, prefixes indicating compass directions are separated from the name, which allows to retrieve documents with more specific location names if a more general one was used in the query. Thus, a search for “*Deutschland*”/ ‘*Germany*’ will also return documents containing the phrase “*Norddeutschland*”/ ‘*Northern Germany*’ (exception: “*Südafrika*”/ ‘*South Africa*’).

We performed first experiments with *semantic representation matching* for GIR at GeoCLEF 2005 [1]. GIR-InSicht is derived from the deep QA system InSicht [2] and matches reduced semantic networks (SNs) of the topic description (or topic title) to the SNs of sentences from the document collection. This process is quite strict and proceeds sentence by sentence.⁴ Before matching starts, the query SN is allowed to be split in parts at specific semantic relations, e.g. at a LOC relation (location of a situation or object) of the MultiNet formalism (multilayered extended semantic networks; [3]), to increase recall while not losing too much precision.

For GeoCLEF 2007, *query decomposition* was implemented, i.e. a query can be decomposed into two queries. First, a geographic subquery about the geographic part of the original query is derived and answered by the QA system InSicht. These geographic answers are integrated into the original query on the SN level (thereby avoiding the complicated or problematic integration on the surface level) yielding one or more revised queries. For example, the query ‘*Whiskey production on the Scottish Islands*’ (57-GC) leads to the geographic subquery ‘*Name Scottish islands*’. GIR-InSicht also decomposes the alternative query SNs derived by inferential query expansion. In the above example, this results in the subquery ‘*Name islands in Scotland*’. InSicht answers the subqueries on the SNs of the GeoCLEF document collection and the German Wikipedia. For the above subqueries, it correctly delivered islands like ‘*Iona*’ and ‘*Islay*’, which in turn lead to revised query SNs which can be paraphrased as ‘*Whiskey production*

³ <http://www.auswaertiges-amt.de/diplo/de/Infoservice/Terminologie/Staatennamen.pdf>

⁴ But documents can also be found if the information is distributed across several sentences because a coreference resolver processed the SN representation for all documents.

on Iona’ and ‘Whiskey production on Islay’. Note that the revised queries are processed only as alternatives to the original query.

Another decomposition strategy produces questions aiming at meronymy knowledge based on the geographic type of a location, e.g. for a country *C* in the original query a subquery like “Name cities in *C*” is generated, whose results are integrated into the original query SN yielding several revised queries. This strategy led to interesting questions like ‘Which country/region/city is located in the Himalaya?’ (GC-69). In total, both decomposition strategies led to 80 different subqueries for the 25 topics. After the title and description of a topic have been processed independently, GIR-InSicht combines the results. If a document occurs in the title results and the description results, the highest score was taken for the combination.

The semantic matching approach is completely independent of the main approach in GIRSA. Some of the functionality of the main approach is also realized in the matching approach, e.g. some of the location indicator classes described above are also exploited in GIR-InSicht (adjectives; demononyms for regions and countries). These location indicators are not normalized, but the query SN is extended by many alternative SNs that are in part derived by symbolic inference rules using the semantic knowledge about location indicators. In contrast, the main approach exploits this information on the level of terms.

There has been little research on the role of normalization of location names, inferring locations from textual clues, and applying QA to GIR. Nagel [4] describes the manual construction of a place name ontology containing 17,000 geographic entities as a prerequisite for analyzing German sentences. He states that in German, toponyms have a simple inflectional morphology, but a complex (idiosyncratic) derivational morphology. Buscaldi, Rosso et al. [5] investigate the semi-automatic creation of a geographic ontology, using resources like Wikipedia, WordNet, and gazetteers. Li, Wang et al. [6] introduce the concept of implicit locations, i.e. locations which are not explicitly mentioned in a text. The only case explored are locations that are closely related to other locations. Our own previous work on GIR includes experiments with documents and queries represented as SNs [1], and experiments dealing with linguistic phenomena, such as identifying metonymic location names to increase precision in GIR [7]. Metonymy recognition was not included in GIRSA because we focused on investigating means to increase recall.

3 Experimental Setup

GIRSA is evaluated on the data from GeoCLEF 2007, containing 25 topics with a title, a short description, and a narrative part. As for previous GIR experiments on GeoCLEF data [1], documents were indexed with a database management system supporting standard relevance ranking (*tf-idf* IR model). Documents are preprocessed as follows to produce different indexes:

1. S: As in traditional IR, all words in the document text (including location names) are stemmed, using a snowball stemmer for German.

Table 2. Frequencies of selected location indicator classes

Class	# Documents	# Locations	# Unique locations
demonym	23379	39508	354
location abbreviation	33697	63223	248
location adjective	211013	751475	2100
location name	274218	2168988	16840
all	284058	3023194	17935

Table 3. Results for different retrieval experiments on German GeoCLEF 2007 data

Run ID	Parameters			Results						
	query	language	index	fields	rel	ret	MAP	P@5	P@10	P@20
FUHtd1de	DE	S	TD	597	0.119	0.280	0.256	0.194		
FUHtd2de	DE	SL	TD	707	0.191	0.288	0.264	0.254		
FUHtd3de	DE	SLD	TD	677	0.190	0.272	0.276	0.260		
FUHtdn4de	DE	SL	TDN	722	0.236	0.328	0.288	0.272		
FUHtdn5de	DE	SLD	TDN	717	0.258	0.336	0.328	0.288		
FUHtd6de	DE	SLD/O	TD	680	0.196	0.280	0.280	0.260		
GIR-InSicht	DE	O	TD	52	0.067	0.104	0.096	0.080		
FUHtd1en	EN	S	TD	490	0.114	0.216	0.188	0.162		
FUHtd2en	EN	SL	TD	588	0.146	0.272	0.220	0.196		
FUHtd3en	EN	SLD	TD	580	0.145	0.224	0.180	0.156		
FUHtdn4en	EN	SL	TDN	622	0.209	0.352	0.284	0.246		
FUHtdn5en	EN	SLD	TDN	619	0.188	0.272	0.256	0.208		

2. SL: Location indicators are identified and normalized to a base form of a location name.
3. SLD: In addition, document words are decomposed. German decomposing follows the frequency-based approach described in [8].
4. O: Documents and queries are represented as SNs and GIR is seen as a form of QA.

Typical location indicator classes were selected for normalization in documents and queries. Their frequencies are shown in Table 2.

Queries and documents are processed in the same way. The title and short description were used for creating a query. GeoCLEF topics contain a narrative part describing documents which are to be assessed as relevant. Instead of employing a large gazetteer containing location names as a knowledge base for query expansion, additional location names were extracted from the narrative part of the topic.

For the bilingual (English-German) experiments, the queries were translated using the Prompt web service for machine translation.⁵ Query processing then follows the setup for monolingual German experiments.

⁵ <http://www.e-prompt.com/>

Values of three parameters were changed in the experiments, namely the query language (German: DE; English: EN), the index type (stemming only: S; identification of locations, not stemmed: SL; decomposition of German compounds: SLD; based on SNs: O; hybrid: SLD/O), and the query fields used (combinations of title T, description D, and locations from narrative N). Parameters and results for the GIR experiments are shown in Table 3. The table shows relevant and retrieved documents (rel_ret), MAP and precision at five, ten, and twenty documents. In total, 904 documents were assessed as relevant for the 25 topics. For the run FUHtd6de, results from GIR-InSicht were merged with results from the experiment FUHtd3de in a straightforward way, using the maximum score. (Run IDs indicate which parameters and topic language were used.)

4 Results and Discussion

Identifying and indexing normalized location indicators, compounding, and adding location names from the narrative part improves performance significantly (paired Student's t-test, $P=0.0008$), i.e. another 120 relevant documents are found and MAP is increased from 0.119 (FUHtd1de) to 0.258 for FUHtdn5de.

Decompounding German nouns seems to have different effects on precision and recall (FUHtd2de vs. FUHtd3de and FUHtdn4de vs. FUHtdn5de). More relevant documents are retrieved without decompounding, but initial precision is higher with decompounding.

The topic '*Deaths caused by avalanches occurring in Europe, but not in the Alps*' (55-GC) contains a negation in the topic title and description. However, adding location names from the narrative part of the topic ('*Scotland, Norway, Iceland*') did not notably improve precision for this topic (0.005 MAP in FUHtd3de vs. 0.013 MAP in FUHtdn5de).

A small analysis of results found by GIR-InSicht in comparison with the main GIR system revealed that GIR-InSicht retrieved documents for ten topics and returned relevant documents for seven topics. This approach contributes three additional relevant documents to the combination (FUHtd6de).

For the topic '*Crime near St. Andrews*' (52-GC), zero relevant documents were retrieved in all experiments. Several topics had a high negative difference to the median average precision, i.e. their performance was lower. These topics include "*Schäden durch sauren Regen in Nordeuropa*" ('*Damage from acid rain in northern Europe*', 54-GC), "*Beratungen der Andengemeinschaft*" ('*Meetings of the Andean Community of Nations*', 59-GC), and "*Todesfälle im Himalaya*" ('*Death on the Himalaya*', 69-GC). The following causes for the comparatively low performance were identified:

- The German decompounding was problematic with respect to location indicators, i.e. location indicator normalization was not applied to the constituents of German compounds (e.g. "*Andengemeinschaft*" is correctly split into "*Anden*"/'*Andes*' and "*Gemeinschaft*"/'*community*', but "*Anden*" is not identified as a location name for topic 59-GC).

- Several terms were incorrectly stemmed, although they were base forms or proper nouns (e.g. “*Regen*”/‘rain’ → “*reg*” and “*Anden*”/‘Andes’ → “*and*” for topics 54-GC and 59-GC, respectively).
- Compounding led in some cases to terms with a very high frequency, causing a thematic shift in the retrieved documents (e.g. “*Todesfälle*”/‘cases of death’ was split into “*Tod*”/‘death’ and “*Fall*”/‘case’ for topics 55-GC and 69-GC).
- In several cases, a focused query expansion might have improved performance, i.e. “*Scandinavia*” may have been a good term for query expansion in topic 54-GC, but GIRSA’s main approach did not use query expansion for GeoCLEF 2007.

Results for the bilingual (English-German) experiments are generally lower. As for German, all other experiments outperform the baseline (0.114 MAP). The best performance is achieved by an experiment using topic title, description, and location names from the narrative (0.209 MAP). In comparison with results for the monolingual German experiments, the performance drop lies between 4.2% (first experiment) and 27.1% (fifth experiment).

The narrative part of a topic contains a detailed description about which documents are to be assessed as relevant (and which not), including additional location names. Extracting location names from the narrative (instead of looking up additional location names in large gazetteers) and adding them to the query notably improves performance. This result is seemingly in contrast to some results from GeoCLEF 2006, where it was found that additional query terms (from gazetteers) degrade performance. A possible explanation is that in this experiment, only a few location names were added (3.16 location names on average for 15 of the 25 topics with a maximum of 13 additional location names). When using a gazetteer, one has to decide which terms are the most useful ones in query expansion. If this decision is based on the importance of a location, a semantic shift in the results may occur, which degrades performance. In contrast, selecting terms from the narrative part increases the chance to expand a query with relevant terms only.

5 Conclusion and Outlook

In GIRSA, location indicators were introduced as text segments from which location names can be inferred. Results of the GIR experiments show that MAP is higher when using location indicators instead of geographic proper nouns to represent the geographic scope of a document. This broader approach to identify the geographic scope of a document benefits system performance because proper nouns or location names do not alone imply the geographic scope of a document.

The hybrid approach for GIR proved successful, and even a few additional relevant documents were found in the combined run. As GIR-InSicht originates from a deep (read: semantic) QA approach, it returns documents with a high initial precision, which may prove useful in combination with a geographic blind feedback strategy. GIR-InSicht performs worse than the IR baseline, because

only 102 documents were retrieved for 10 of the 25 topics. However, more than half (56 documents) turned out to be relevant.

Several improvements are planned for GIRSA. These include using estimates for the importance (weight) of different location indicators, possibly depending on the context (e.g. *'Danish coast' → 'Denmark'*, but *'German shepherd' ↗ 'Germany'*), and augmenting the location name identification with a part-of-speech tagger and a named entity recognizer.

Furthermore, the QA methods provide a useful mapping from natural language questions to gazetteer entry points. For example, the expression *'Scottish Islands'* is typically not a name of a gazetteer entry, while the geographic subquery answers *'Iona'* and *'Islay'* typically are. In the future, a tighter coupling between the QA and IR components is planned, exploiting these subquery answers in the IR methods of GIRSA. (Note that this reverses the standard order of processing known from QA: In GIRSA, QA methods are employed to improve performance before subsequent IR phases.) Finally, we plan to investigate the combination of means to increase precision (e.g. recognizing metonymic location names) with means to increase recall (e.g. recognizing and normalizing location indicators).

References

1. 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., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 977–986. Springer, Heidelberg (2006)
2. Hartrumpf, S., Leveling, J.: Interpretation and normalization of temporal expressions for question answering. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 432–439. Springer, Heidelberg (2007)
3. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Berlin (2006)
4. Nagel, S.: An ontology of German place names. Corela – Cognition, Représentation, Langage – Le traitement lexicographique des noms propres (2005)
5. Buscaldi, D., Rosso, P., Garcia, P.P.: Inferring geographical ontologies from multiple resources for geographical information retrieval. In: Proceedings of GIR 2006, Seattle, USA, pp. 52–55 (2006)
6. Li, Z., Wang, C., Xie, X., Wang, X., Ma, W.Y.: Indexing implicit locations for geographical information retrieval. In: Proceedings GIR 2006, Seattle, USA, pp. 68–70 (2006)
7. Leveling, J., Hartrumpf, S.: On metonymy recognition for GIR. In: Proceedings of GIR 2006, Seattle, USA, pp. 9–13 (2006)
8. Chen, A.: Cross-language retrieval experiments at CLEF 2002. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2002. LNCS, vol. 2785, pp. 28–48. Springer, Heidelberg (2003)

GeoParsing Web Queries

Rocio Guillén

California State University San Marcos, San Marcos, CA 92096, USA
rguillen@csusm.edu

Abstract. In this paper we present preliminary results for a pattern-based approach to parse web-based queries. The approach is designed to identify and categorize queries that include a geographical reference. Due to the ungrammaticality, multilinguality and ambiguity of the language in the 800,000 web-based queries in the collection, we started by building a list of all the different words in the queries, similar to creating an index. Next, a lookup of the words was done in a list of countries to identify potential locations. Because many locations were missed, we further analyzed the queries looking for spatial prepositions and syntactic cues. Queries were processed by combining search in gazetteers with a set of patterns. Categorization was also based on patterns. Results were low in terms of recall and precision mainly because the set of patterns is incomplete. Further statistical analysis and application of machine learning techniques is likely to improve performance. Error analysis of the results is discussed in detail.

1 Introduction

Increase in the usage of the web has motivated researchers in Information Retrieval to study other dimensions to the data that help separate useful resources from less useful ones in an extremely heterogeneous environment like the web ([1]). It has been found that web users often issue queries that include geographically specific information related to accommodation, entertainment, medical services, real estate, and cultural events. A study of a log of the Excite search engine showed that one fifth of all queries were geographical, as determined by the presence of a geographical reference such as the name of a place ([2]) or a special relationship to qualify the name of the place.

This paper addresses the problem of identifying geographic entities and spatial relationships in a log of 800,000 web queries, discovering the local and global geographic context of the query, and categorizing the queries according to their information content.

We first define how to determine whether a query includes a geographic entity by identifying patterns using gazetteers. We then describe how candidate geographical queries are further processed to extract the specific location. Finally, we discuss the limitations of the approach by analyzing errors based on a comparison between our results with the results produced by the task evaluators.

2 Approach

Geoparsing is the process of recognizing geographic context ([3]). The first step involves extracting geographic entities from texts and distinguishing them from other entities such as names of people or organizations, and events. In natural language processing this is referred to as Named Entity Recognition (NER) and is central to other text processing applications such as information extraction (IE) and information retrieval (IR).

Our focus is on geoparsing a log of the MSN search engine, one of the tasks in GeoCLEF2007 ([4]). The query parsing task consisted of parsing queries to recognize and extract georeferences, which included geographical relationships such as *where*, *geospatial relation*, and *type* of geographical query ([5]).

Traditional Information Extraction (IE) has involved manual processing in the form of rules or tagging training examples where the user is required to specify the potential relationships of interest ([6]). The main focus of IE has been on extracting information from homogeneous corpora such as newswire stories. Hence, traditional IE systems rely on linguistic techniques applied to the domain of interest, such as syntactic parsers and named-entity recognizers. The problem of extracting information from Web-based corpora presents different challenges. The use of name-entity recognizers and syntactic parsers encounters problems when applied to heterogeneous text found on the Web, and web-based queries are no exception.

Current work on query processing for retrieving geographic information on the Web has been done by Chen *et. al* ([7]). Their approach requires a combination of text and spatial data techniques for usage in geographic web search engines. A query to such an engine consists of keywords and the geographic area the user is interested in (i.e., query footprint). Our approach combines information extraction and patterns.

Due to the ungrammaticality, multilinguality and ambiguity of the language in the 800,000 web-based queries in the collection, we started by building a list of all the different words, similar to creating an index, excluding stopwords. Next, a lookup of the words was done in a list of countries, main cities and states to identify potential locations. The list was created from the GEOnet Names Server database ([8]). One problem is multiword georeferences. Because many locations were missed, we selected those queries where spatial prepositions such as “in”, “near” and syntactic cues, such as “lake”, “street”, “hotel”, “accommodation”, were present. We have considered these as good heuristics for recognizing multiword expressions as georeferences and to create pattern-based rules to further process potential candidates.

Extraction of geographical information such as latitude and longitude was done as follows. We created a new list of words identified as potential geographic references. Latitude and longitude information was looked up in the GNS database. A problem that we found is related to ambiguity since a geographic reference may refer to a city, state, park, and the same geographic entity may be in different continents, countries, states, and cities.

Finally, categorization was done using patterns. If the only information available was the name of a place, the query was categorized as of type “Map”. If words such as “college”, “airport”, “studio” were present, the query was categorized as of type “Yellow Page”. If the query included words such as “flight”, “survey”, “company”, the query was categorized as of type “Information”.

3 System Description

The system architecture has the following modules: 1) Preprocessor, 2) Georeference Identification, 3) Categorization and 4) Geocoding. We wrote programs in C for all the components; they run in a Linux environment.

3.1 Preprocessor

The Preprocessor module takes the query log as input to remove XML tags and create one line per query with the contents of the `< QUERYNO >` tag and the contents of the `< QUERY >` tag. Additionally it creates a file with all the words (except stopwords in English) in the log that it is sorted to eliminate duplicates. A file (locations) with countries, capitals, and main cities is also created from different gazetteers available on the web.

3.2 Georeference Identification

This module compares the new query file and the locations file looking up for a match. If a match is not found the program checks for patterns. It outputs two files: 1) a set of queries (local-query file), where potential locations or patterns were identified for further processing, and 2) a non-local-query file. We analyzed both files manually to add information to the locations file and the patterns.

3.3 Categorization

We examined the training file manually to derive rules for determining the type of information contained in a query. The categorization component takes the local-query file and applies the rules to automatically tag each query. If a pattern is missing no tag is assigned.

3.4 Geocoding

Geocoding was done by looking up the information in the “where” component of each query in a subset of the GNS database. The final step was to merge the local-query file and the non-local-query file.

4 Results

Six teams participated in the query parsing task. The highest precision score was 0.625, the highest recall was 0.566 and the highest F1 score was 0.488. The

lowest precision score was 0.096, the lowest recall was 0.038, and the lowest F1 score was 0.088. Our results in terms of precision, recall and F1 were 0.201, 0.197, and 0.199, respectively. Three systems achieved better scores than ours. A detailed description of the results for all the participants is presented in the task evaluators report ([4],[5]).

5 Error Analysis

Our initial approach is limited because we did not identify all the patterns and the parser is not complete, which hindered its performance. In this section, we discuss two types of errors: 1) Those related to geographical references *where*, *geo-relation*, and *local* query; and 2) those related to *what-type* of geographical query.

5.1 Where, Geo-relation and Local Mismatches

A problem with the *where* component was making the assumption that it was sufficient to recognize and extract the location from the query, when it was required to include additional information such as the country. For instance, the *where* component in the query “caverns in Alabama” was “Alabama, United States” instead of “Alabama”. Geographical locations without country accounted for 43% of the total errors, incomplete geographical locations without country accounted for 15% of errors, false negatives accounted for 13% of errors, false positives accounted for 11% of errors, and the rest are errors in extracting non-geographical references.

Errors in extracting the correct geo-relations are as follows: false negatives 47% of the total errors, false positives 23%, incomplete geo-relations 18% (e.g., “of” instead of “north_of”), the rest are errors in extracting the correct geographical relation. For instance take the query “fishing in southwest Idaho”, the candidate relation was “in” instead of “southwest_of”. When analyzing results for the *local* component, we found that 48% of the total errors were false positives and 52% were false negatives. Out of 400 queries 19.5% were misclassified.

Examining the results produced by the evaluators, we found some apparent contradictions related to local vs. nonlocal queries. For instance the query in Spanish “*peru autos usados*” transliterated into English as “peru used cars” was not classified as a local query even when “*peru*” is a country. However, the query in French “*carrefour jeunesse emploi*” transliterated into English as “carrefour youth employment” was classified as a local query in Canada. This would have an impact on the evaluation of the results submitted.

5.2 What-Type Content

Categorization of type of local queries into *information*, *map* and *yellow page* represented a challenge. In many cases there is not a clear definition of each category. Miscategorization of queries (222) made up 55 percent of cases distributed as shown in Table 1.

Table 1. Distribution of Miscategorized Queries

Type	False Positives	False Negatives
<i>Yellow Page</i>	7	89
<i>Map</i>	14	55
<i>Information</i>	1	56

6 Conclusion

Parsing and categorization of web-based queries is a challenging task because of the nature of the data. It is syntactically, semantically and pragmatically ambiguous. Processing multilingual ungrammatical data requires more resources such as bilingual dictionaries and multilingual gazetteers to help disambiguate parts-of-speech from a linguistic point of view to answer the “where” component more accurately. The ambiguity of the content information requires of a clearer definition of the categories to which the queries can be assigned to better determine the “what” component. Therefore, a pattern-based approach is limited as shown by the results. However, it has provided a better understanding of the difficulty of geoparsing and geocoding real web-based queries. Further investigation and application of classical and statistical language processing techniques is needed to improve the performance of the approach presented. Current work is focused on inferring a grammar and eventually a language model that would improve the performance of the approach presented.

References

1. Gravano, L., Hatzivassilogiou, V., Lichtenstein, R.: Categorizing Web Queries According to Geographical Locality. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM 2003, pp. 325–333 (2003)
2. Sanderson, M., Kohler, J.: Analyzing Geographic Queries. In: Proceedings Workshop on Geographical Information Retrieval SIGIR (2004), <http://www.geounizh.ch/~rsp/gir/>
3. Larson, R.R.: Geographic Information Retrieval and Spatial Browsing. In: Smith, L., Gluck, M. (eds.) University of Illinois GIS and Libraries: Patrons, Maps and Spatial Information, pp. 81–124 (1996)
4. Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xie, X.: GeoCLEF2007: the CLEF2007 Cross-Language Geographic Information Retrieval Track Overview. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop 2007, Budapest, Hungary (2007)
5. Li, Z., Wang, C., Xie, X., Ma, W.: Query Parsing Task for GeoCLEF2007 Report. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop 2007, Budapest, Hungary (2007)
6. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open Information Extraction from the Web. In: Proceedings Twentieth International Joint Conference on Artificial Intelligence 2007, pp. 2670–2676 (2007)
7. Chen, Y., Suel, T., Markowetz, A.: Efficient Query Processing in Geographic Web Search Engines. In: Proceedings SIGMOD 2006, pp. 277–288 (2006)
8. <http://earth-info.nga.mil/gns/html/index.html>

MIRACLE at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information

Sara Lana-Serrano^{1,3}, Julio Villena-Román^{2,3},
José Carlos González-Cristóbal^{1,3}, and José Miguel Goñi-Menoyo¹

¹ Universidad Politécnica de Madrid

² Universidad Carlos III de Madrid

³ DAEDALUS - Data, Decisions and Language, S.A.

slana@diatel.upm.es, jvillena@it.uc3m.es,
josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es

Abstract. This paper describes the participation of MIRACLE research consortium at the Query Parsing task of GeoCLEF 2007. Our system is composed of three main modules. The first one is the Named Geo-entity Identifier, whose objective is to perform the geo-entity identification and tagging, i.e., to extract the “where” component of the geographical query, if there is any. Then, the Query Analyzer parses this tagged query to identify the “what” and “geo-relation” components by means of a rule-based grammar. Finally, a two-level multiclassifier first decides whether the query is indeed a geographical query and, should it be positive, then determines the query type according to the type of information that the user is supposed to be looking for: map, yellow page or information.

Keywords: Linguistic Engineering, classification, geographical IR, geographical entity recognition, gazetteer, Geonames, tagging, query classifier, WordNet.

1 Introduction

MIRACLE team is a research consortium formed by research groups of three different Spanish universities (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a private company founded as a spin-off of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in most tracks and tasks, including the main bilingual, monolingual and cross lingual tasks [1] as well as in ImageCLEF, WebCLEF, GeoCLEF [2] [3] and Question Answering tracks.

This paper describes the MIRACLE participation [4] at the Query Parsing task of GeoCLEF 2007 [5]. In the following sections, we will first give an overview of the architecture of our system. Afterwards we will elaborate on the different modules. Finally, the results will be presented and analyzed.

2 System Description

The system architecture is shown in Figure 1. Note that our approach consists of three sequential tasks executed by independent modules [4]:

- **Named Geo-entity Identifier:** performs the geo-entity identification and a query expansion with geographical information.
- **Query Analyzer:** identifies the “what” and “geo-relation” components of a geographical query.
- **Query Type Classifier:** determines the type of geographical query.

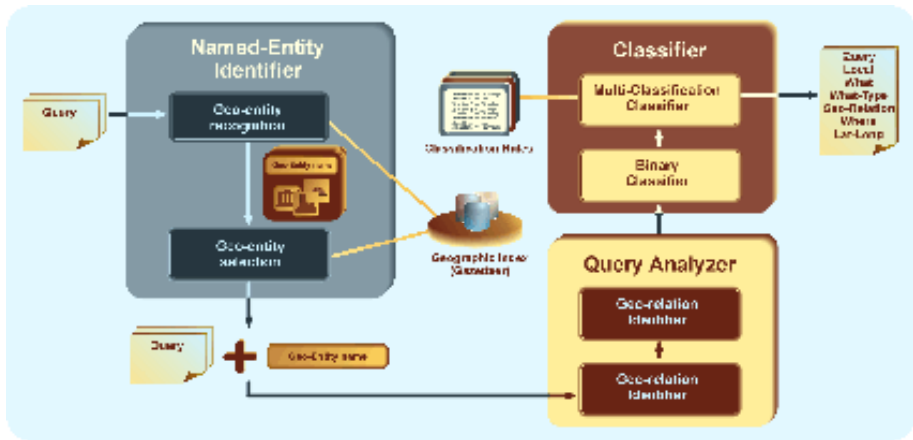


Fig. 1. Overview of the system

2.1 Named Geo-entity Identifier

The objective of this module is to perform the geo-entity identification and tagging, i.e., to extract the “where” component of the query, should there be any. It is composed of two main components: a gazetteer, i.e. a database with geographical resources that constitutes the knowledge base of the system, and a geo-entity parser built on top of it.

Our gazetteer is built up from the Geonames geographical database [6], available free of charge for download under a Creative Commons attribution license. It contains over 8 million geographical names with more than 6.5 million unique features about 2.2 million populated places and 1.8 million alternate names. Those features include a unique identifier, the resource name, alternative names (in other languages), county/region, administrative divisions, country, continent, longitude, latitude, population, elevation and timezone. All features are categorized into one out of 9 feature classes and further subcategorized into one out of 645 feature codes. Geonames integrates geographical data (such as names of places in various languages, elevation or population) from various sources, mainly the Geonet Names Server (GNS) [7] gazetteer of the National Geospatial Intelligence Agency (NGA), the Geographic Names Information System (GNIS) [8] gazetteer of the U.S. Geographic Survey, the GTOPO30 [9]

digital elevation model for the world developed by United States Geological Survey (USGS) and Wikipedia, among others.

For our purposes, all data was loaded and indexed in a MySQL database, although not all fields (such as time zone or elevation) are used: the relevant fields are UFI (unique identifier), NAME_ASCII (name), NAME_ALTERNATE (alternate names), COUNTRY, ADM1 and ADM2 (administrative region where the entity is located), FEATURE_CLASS, FEATURE_TYPE, POPULATION, LATITUDE and LONGITUDE. To simplify the query processing, each row is complemented with the expansion of country codes (ES→Spain) and/or state codes (NC→North Carolina) – when applicable. The final database uses 865KB.

The geo-entity parser carries out the following tasks:

- **Geo-entity recognition:** identifies named geo-entities [3] using the information stored in the gazetteer, looking for candidate named entities matching any substring of one or more words [10] included in the query and not included in a stopword (or stop-phrase) list [11].
The stopword list is mainly automatically built by extracting those words that are both common nouns and also georeference entities, assuming that the user is asking for the common noun sense (for example, “Aguilera” –for “Christina Aguilera”– or “tanga” – “thong”). Specifically we have used lexicons for English, Spanish, French, Italian, Portuguese and German, and have selected words that appear at least with a certain frequency in the query collection. The stopword list currently contains 1,712 entries.
- **Geo-entity selection:** The selected named geo-entity will be the one with the longest number of matching words and, if the same, the one with higher score. The score is computed according to the type of geographic resource (country, region, county, city...) and its population, as shown in the following table.

Table 1. Entity score

Feature type	Code	Score
Capital and other big cities	PPLA, PPLC, PPLG	Population+100,000,000
Political entities	PCL, PCLD, PCLF, PCLI, PCLIX, PCLS	Population+10,000,000
Countries	A	Population+1,000,000
Other cities	PP, STLMT	Population+100,000
Other	*	Population
For all cities, if country/state name/code is also in the query	PP, STLMT	Score += 100,000,000

Those values were arbitrarily chosen after different manual executions and subsequent analysis.

- **Query tagging:** expands the query with information about the selected entity: name, country, longitude, latitude, and type of geographic resource.

The output of this module is the list of queries in which a possible named geo-entity has been detected, along with their complete tagging. Table 2 shows an example of a possible output.

Table 2. Example of tagged geo-entities

Query	score	ufi	entity	state (code)	country (code)	latitude	longitude	feature_class	feature_type	
airport {{alicante}}	car rental week	2693959	2521976	Alicante	Spain (ES)	38.5	-0.5	A	ADM2	
bedroom apartments for sale in {{bulgaria}}		10000000	732800	Bulgaria (BG)	43.0	25.0	A	PCLI		
hotels in {{south lake tahoe}}		123925	5397664	South Lake Tahoe	California (CA)	United States (US)	38.93	-119.98	P	PPL
Helicopter flight training in southwest {{florida}}		100100000	4920378	Florida	Indiana (IN)	United States (US)	40.16	-85.71	P	PPL

Note that the geo-entity is specifically marked in the original query, enclosed between double curly brackets, to help the following module to identify the rest of the components of the geographical query.

2.2 Query Analyzer

This module parses each previously tagged query to identify the “what” and “geo-relation” components of a geographical query, sorting out the named geo-entity detected by the previous module, enclosed between curly brackets. It, in turn, consists of two subsystems:

- **Geo-relation identifier:** identifies and qualifies spatial relationships using rule-based regular expressions. Its output is the input list of queries expanded with information related to the identified “geo-relation”.

Table 3 shows the output of this module for the previous examples.

Table 3. Geo-relation expansion

<i>Query</i>	<i>geo-relation</i>	<i>entity</i>	<i>state</i>	<i>country</i>	<i>country (code)</i>	<i>latitude</i>	<i>longitude</i>	<i>feature_class</i>	<i>feature_type</i>
airport {{alicante}}	car rental week	NONE	Alicante	Spain	ES	38.5	-0.5	A	ADM2
bedroom apartments for sale #@#IN#@# {{bulgaria}}	IN	Bulgaria	Bulgaria	BG	43.0	25.0	A	PCLI	
hotels #@#IN#@# {{south lake tahoe}}	IN	South Lake Tahoe	California	United States	US	38.93	-119.98	P	PPL
helicopter flight training in #@#SOUTH_WEST_OF#@# {{florida}}	SOUTH_WEST_OF	Florida	Indiana	United States	US	40.16	-85.71	P	PPL

Note that the geo-relation is also marked in the original query.

- **Concept identifier:** analyses the output of the previous step and extracts the “what” component of a geographical query applying manually defined grammar rules based on the identified “where” and “geo-relation” components.

2.3 Query Type Classifier

Finally, the last step is to decide whether the query is indeed a geographical query and, should it be positive, to determine the type of query, according to the type of information that the user is supposed to be looking for:

- **Map type:** users are looking for natural points of interest, such as rivers, beaches, mountains, monuments, etc.
- **Yellow page type:** businesses/organizations, like hotels, restaurants, hospitals, etc.
- **Information type:** users are looking for text information (news, articles, blogs).

The process is carried out by a two level classifier [4]:

1. **First level:** a binary classifier to determine whether a query is a geographical or a non-geographical query. This simple classifier is based on the assumption that a query is geographical if the “where” component is not empty.
2. **Second level:** a multi-classification rule-based classifier to determine the type of geographical query. The multi-classifier treats the tagged queries as a lexicon of semantically related terms (words, multi-words and query parts).

The classification algorithm applies a knowledge base that consists on a set of manually defined grammar rules, including nouns and grammatically related part-of-speech categories as well as the type of geographical resource. The different valid lemmas are unified using Wordnet synsets [4].

3 Results

For the evaluation, multiple human editors labeled 500 queries that were chosen to represent the whole query set. Then all the submitted results were manually compared to those queries following a strict criterion where a match should have all fields correct. Table 4 shows the evaluation results of our submission, using the well-known evaluation measures of precision, recall and F1-score.

According to the task organizers [5], our submission achieved the best performance (F1-score) out of the 6 submissions of this year, which was satisfying, given our hard work. Other groups used similar approaches, but we think that the coverage of our

Table 4. Overall results

Precision ⁽¹⁾	Recall ⁽²⁾	F1-score ⁽³⁾
0.428	0.566	0.488

$$\text{precision} = \frac{\text{correctly\_tagged\_queries}}{\text{all\_tagged\_queries}} \tag{1}$$

$$\text{recall} = \frac{\text{correctly\_tagged\_queries}}{\text{all\_relevant\_queries}} \tag{2}$$

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

gazetteer, an adequate stopword list, the algorithm for geo-entity selection and the precision of the query classifier let us make the difference with respect to other systems.

In addition, as participants in the task were provided with the evaluation data set, we have further evaluated our submission to separately study the results for each component of the geographical queries and also analyze the level-by-level performance of the final classifier.

Table 5 shows the individual analysis of the classifier per each field. The first-level classifier (LOCAL) achieves a precision of 75.40%, but the second-level classifier reduces this value to 56.20% for the WHAT-TYPE feature. According to a strict evaluation criterion, this would be the precision of the overall experiment.

Table 5. Individual analysis per field

	LOCAL		WHAT		WHAT-TYPE		WHERE		ALL	
	Total	%	Total	%	Total	%	Total	%	Total	%
All topics	377	75.40	323	64.60	281	56.20	321	64.20	259	51.80
Well-classified	377	100.00	323	85.67	281	74.53	321	85.15	259	68.70

However, if evaluated only over well-classified (geographical/non geographical) queries, the precision arises to 74.53% for the same feature. This great improvement shows that the precision of the system highly depends on the correct classification of the query and the first-level classifier turns out to be one of the key components of the system. The confusion matrix for this classifier is shown in Table 6, which shows that the precision is 73%. The conclusion for future participations is that more effort should be invested on improving this classifier to increase the overall performance.

Table 6. Confusion matrix for the binary classifier

	LOCAL		
	LOCAL YES	LOCAL NO	
ASSIGNED YES	297	111	Precision⁽¹⁾
ASSIGNED NO	12	80	Recall⁽²⁾
			Accuracy⁽³⁾
			0.73
			0.96
			0.75

$$^{(1)} \text{ precision} = \frac{TP}{TP + FP}$$

$$^{(2)} \text{ recall} = \frac{TP}{TP + FN}$$

$$^{(3)} \text{ accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 7 shows the same evaluation for the multiclassifier, but individualized per class and calculated over all topics. The lowest precision corresponds to “Yellow Page” queries. The explanation is that our gazetteer lacks that type of information such as names of hotels, hospitals, shopping centers, etc. This issue will be solved for future participations.

The following Table 8 shows the same evaluation per class, but calculated only over topics which are correctly classified by the first-level binary classifier. It is interesting to observe an increase in precision for all types of queries, but the relative distribution remains the same. As in the previous table, the lowest recall corresponds to “Map” queries. The difficulty to classify, parse and execute these queries may explain this fact.

Table 7. Evaluation of the multiclassifier, per class, all topics

Type	Precision	Recall	Accuracy
Yellow Page	0.43	0.95	0.61
Map	0.74	0.52	0.89
Information	0.93	0.20	0.88

Table 8. Evaluation of the multiclassifier, per class, correctly-classified topics

Type	Precision	Recall	Accuracy
Yellow Page	0.61	0.99	0.75
Map	0.92	0.55	0.89
Information	1.00	0.21	0.86

Last, we have to express some disagreements with the evaluation data provided by the organizers. Although some issues may be actual errors, most are due to the complexity and ambiguity of the queries. Table 9 shows some examples of queries that have been classified as geographical by our system but have been evaluated as false-positives. In fact, we think that it would be almost impossible to reach a complete agreement in the parsing or classification for every case among different human editors. The conclusion to be drawn from this is that the task to analyze and classify queries is very hard without a previous contact and without the possibility of interaction and feedback with the user.

Table 9. Some examples of ambiguities

QueryNo	Query	Extracted “where”	Why not?
113501	calabria chat	calabria, Italy	chat rooms about the region of Calabria?
443245	Machida	machida, Japan	Hiroko Machida (actress), Kumi Machida (artist) or the city of Machida?
486273	montserrat reporter	montserrat, Montserrat	online newspaper or reporters in Montserrat?

4 Conclusions and Future Work

According to a strict evaluation criterion where a match should have all fields correct, our system reaches a precision value of 42.8% and a recall of 56.6% and our submission is ranked 1st out of 6 participants in the task.

However, a detailed evaluation of the confusion matrixes reveals that some extra effort must be invested in “user-oriented” disambiguation techniques to improve the first level binary classifier for detecting geographical queries, as it is a key component to eliminate many false-positives.

In addition, the analysis of the confusion matrixes for the multiclassifier that are calculated over the topics correctly classified by the first level classifier shows that the probability that a geographical query is classified as “Yellow Page” is very high. This could be related to the uneven distribution of topics (almost 50% of the geographical queries belong to this class). In addition, “Information” type queries have a very low recall. These combined facts point out that the classification rules have not been able to establish a difference between both classes. We will focus on this issue in future participations.

Acknowledgements. This work has been partially supported 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; and by the Madrid’s R&D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

1. Goñi-Menoyo, J.M., González-Cristóbal, J.C., Villena-Román, J.: MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. 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. Lana-Serrano, S., Goñi-Menoyo, J.M., González-Cristóbal, J.C.: MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR. 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. 920–923. Springer, Heidelberg (2006)
3. Goñi-Menoyo, J.M., González-Cristóbal, J.C., Lana-Serrano, S., Martínez-González, A.: MIRACLE’s Ad-Hoc and Geographic IR approaches for CLEF 2006. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Lana-Serrano, S., Villena-Román, J., Goñi-Menoyo, J.M.: MIRACLE at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information. In: Nardi, A., Peters, C. (eds.) Working Notes of the Cross Language Evaluation Forum (CLEF) 2007 Workshop, Budapest, Hungary (2007)
5. Zhisheng, L., Chong, W., Xing, X., Wei-Ying, M.: Query Parsing Task for GeoCLEF2007 Report. In: Nardi, A., Peters, C. (eds.) Working Notes of the Cross Language Evaluation Forum (CLEF) 2007 Workshop, Budapest, Hungary (2007)
6. Geonames geographical database, <http://www.geonames.org>
7. U.S. National Geospatial Intelligence Agency, <http://www.nga.mil>
8. U.S. Geological Survey, <http://www.usgs.gov>
9. Global 30 Arc-Second Elevation Data Set, <http://eros.usgs.gov/products/elevation/gtopo30.html>
10. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, USA (2000)
11. University of Neuchatel. Page of resources for CLEF, <http://www.unine.ch/info/clef>

Relevance Measures Using Geographic Scopes and Types

Geoffrey Andogah and Gosse Bouma

Computational Linguistics Group,
Centre for Language and Cognition Groningen (CLCG),
University of Groningen, Groningen, The Netherlands
`{g.andogah,g.bouma}@rug.nl`

Abstract. This paper proposes two kinds of relevance measures to rank documents by geographic restriction: scope-based and type-based. The non-geographic and geographic relevance scores are combined using a weighted harmonic mean. The proposed relevance measures and weighting schemes are evaluated on GeoCLEF 2007 dataset with encouraging performance over the standard IR performance. The best performance is achieved when the importance of non-geographic relevance scores outweigh the importance of geographic relevance scores.

1 Introduction

Geographic information retrieval (Geo-IR) is an information retrieval (IR) system which aims to retrieve documents according to both non-geographic and geographic relevance. Standard relevance measures such as the classic vector space model (VSM) are not sufficient for geographic ranking. There is need for geographic relevance measures to intuitively rank documents according to geographic restriction. A single weighting scheme is required to combine non-geographic and geographic relevance scores to generate a unified document list.

In this paper two kind of geographic relevance measures are proposed: one based on geographic scope assigned to user query and document, and the other based on query geographic type. The paper also proposes a weighted harmonic mean based algorithm to combine geographic and non-geographic relevance scores.

2 Relevance Measures

This section describes relevance measures proposed to perform Geo-IR, and how the measures are combined to rank documents.

2.1 Non-geographic Relevance Measure

The Apache Lucene IR library is used to perform non-geographic search. Lucene's default relevance measure is derived from the vector space model (VSM). The

Lucene relevance score formula combines several factors to determine the document score for a given query [1]:

$$NonSim(q, d) = \sum_{t \text{ in } q} tf(t \text{ in } d) \cdot idf(t) \cdot bst \cdot lN(t.field \text{ in } d) \quad (1)$$

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 , bst is the field boost set during indexing and $lN(t.field \text{ in } d)$ is the normalization value of a field given the number of terms in the field.

2.2 Geographic Scope Based Relevance Measure

The geographic scope based relevance measure uses geographic scopes assigned to queries and documents to rank documents according to query geographic restrictions similar to schemes proposed in [2]. The geographic scope resolver [3] assigns multiple scopes to a document, and ranks the assigned scopes according to relevance from the highly relevant to the least relevant for each document.

The geographic scopes are limited to six categories: continent scope, continent-directional scope (e.g. western Europe), country scope, country-directional scope (e.g. north-of Netherlands), province¹ scope, and province-directional scope (e.g. south-east-of California).

The geographic scope based relevance measure is formally defined according to:

$$ScopeSim(q, d, s) = \sum \sqrt{wt_{(q,s)}} \times \log(1 + wt_{(d,s)}) \quad (2)$$

where; $wt_{(q,s)}$ is the weight assigned to scope s in query q by the scope resolver and $wt_{(d,s)}$ is the weight assigned to scope s in document d by the scope resolver.

2.3 Geographic Type Based Relevance Measure

The relevance measure based on geographic type uses the geographic feature class and type as defined in the database of geographic features. The measure ranks documents by query feature type restriction. The feature class and type as defined in the Geonames.org² database are used to implement the type-based relevance measure.

The type based relevance measure is defined as:

$$TypeSim(q, d) = \frac{1.0}{\sqrt{1 + \frac{N_{qFClass} - N_{qFType}}{N_{qFClass}}}} \quad (3)$$

where; $N_{qFClass}$ is the number of occurrence of the required query feature class in the document, and N_{qFType} is the number of occurrence of the required query feature type in the document. GeoCLEF 2007 topic 10.2452/56-GC: ‘Lakes with

¹ Province: is the first order administrative division of a country.

² <http://www.geonames.org/export/codes.html>

monsters' is used to illustrate how Eq. 3 is used to compute document relevance. The query feature type 'Lake' belongs to class 'H' (i.e. class of hydrographic features such as river, stream, lake, bay, etc.). Each retrieved document is queried for class 'H' and feature type 'Lake'. The number of occurrence of 'H' (i.e. $N_{qFClass}$) and 'Lake' (i.e. N_{qFType}) in the document is used to compute the document's geographic relevance according to Eq. 3.

2.4 Relevance Measure Combination

This section describes formulae proposed to combine the non-geographic relevance measure and geographic relevance measure discussed in Sec 2.1, 2.2 and 2.3.

Linear Interpolated Combination. The linear interpolated combination is derived as:

$$Sim(q, d) = \lambda_T NonSim(q, d) + \lambda_G GeoSim(q, d) \quad (4)$$

$$\lambda_T + \lambda_G = 1 \quad (5)$$

where; λ_T is the non-geographic interpolation factor and λ_G is the geographic interpolation factor. The non-geographic and geographic scores are normalized to $[0, 1]$ before linearly combining the ranked lists. The $GeoSim(q, d)$ in Eq. 4 is replaced by either Eq. 2 or Eq. 3.

Weighted Harmonic Mean Combination. The weighted harmonic mean combination borrows the classic precision and recall combination formula commonly used to measure the performance of information retrieval (IR) systems [4]. The motivation is to determine the importance of non-geographic relevance relative to geographic relevance, and then use the insight to rank documents by both non-geographic and geographic relevance. The weighted harmonic mean combination is defined as:

$$Sim(q, d) = \frac{(1 + \beta) \times GeoSim(q, d) \times NonSim(q, d)}{\beta \times GeoSim(q, d) + NonSim(q, d)} \quad (6)$$

where; β is the relevance importance factor. The following special cases are derived as a consequence of harmonic mean combination:

1. if $\beta = 1$, an equal importance is attached to both non-geographic and geographic relevance.
2. if $\beta = 0$, no importance is attached to non-geographic relevance.
3. if $\beta = \infty$, no importance is attached to geographic relevance.

The interesting feature of this combination is that an optimal value of β where the best performance is achieved is easily spotted. The $GeoSim(q, d)$ in Eq. 6 is replaced by either Eq. 2 or Eq. 3.

Extended Harmonic Mean Combination. The extended harmonic mean combination linearly adds non-geographic relevance measure (see Eq. 1) to the weighted harmonic mean combination (see Eq. 6) as follows:

$$Sim(q, d) = NonSim(q, d) + \frac{(1 + \beta) \times GeoSim(q, d) \times NonSim(q, d)}{\beta \times GeoSim(q, d) + NonSim(q, d)} \quad (7)$$

The $GeoSim(q, d)$ in Eq. 7 is replaced by either Eq. 2 or Eq. 3.

3 Experiments

The University of Groningen participated in GeoCLEF 2007 [5], and this section evaluates the relevance measures and weighting schemes described in Sec. 2 on GeoCLEF 2007 dataset [6].

3.1 Query Processing

In [7], geographic topics are placed into one of the eight categories, according to the way they depend on a place (e.g. UK, St. Andrew, etc.), geographic subject (e.g. city, river, etc.) or geographic relation (e.g. north of, western, etc.). The GeoCLEF 2007 topics are generated according to a similar classification. In this experiment two groups of topics are distinguished: (1) topics which mention places of interest by name, and are resolvable to a scope (GROUP1), and (2) topics which mention only the geographic subject of interest (GROUP2). Table 1 shows GeoCLEF 2007 topics depicting topic grouping, geographic expansion, query formulation and geographic relevance measure used.

Geographic expansion is performed on members of GROUP1 that lack sufficient geographic information or that provide ambiguous geographic information (GROUP1 topics: 51, 59, 60, 61, 63, 65, 66, 70).

A number of GROUP1 topics (56, 68, 72) can be characterized as "geographic subject with non-geographic restriction" [7], with exception of topic 67 which is

Table 1. Example topic grouping and query formulation

Example GROUP1 Topic	
Topic num	10.2452/65-GC
Topic title	Free elections in Africa
Geo-expansion	Add names of African countries and their capitals
Query	Formulated by content of topic title-desc-narr tags
Geo-relevance	Scope based measure (see Eq.2)
Example GROUP2 Topic	
Topic num	10.2452/68-GC
Topic title	Rivers with floods
Geo-expansion	none
Query	Formulated by content of topic title-desc-narr tags
Geo-relevance	Type based measure (see Eq.3)

more complex. Resolving the geographic scope of such topics to a specific place is non trivial. The most reasonable scope for these topics is geographic subject scope: lake, river, beach, city, etc.

3.2 Results

The result of the experiment reported here is based on query formulation and geographic relevance measures as shown in Table 1: GROUP1 and GROUP2 topics are geographically ranked according to Eq. 2 and Eq. 3 respectively, geographic expansion is applied to GROUP1 topics, and queries for GROUP1 and GROUP2 are formulated by content of all topic tags. All the parameters of ranking formulas are tuned on both GeoCLEF 2006 and 2007 datasets.

Harmonic Mean vs. Linear Interpolated Combination. This sub-section compares harmonic mean combination (see Eq. 6) retrieval performance against linear interpolated combination (see Eq. 4). From Figure 1 and Figure 2 the following observations can be made:

1. the system performs very poorly at $\lambda_T = 0$ and $\beta = 0$ which represents pure geographic retrieval.
2. the system performs poorly at $\lambda_T = 0.5$ and $\beta = 1$ which gives equal importance to geographic retrieval and non-geographic retrieval in comparison to pure non-geographic retrieval at $\lambda_T = 1.0$ and $\beta = \infty$.
3. the best system performance is observed at $\lambda_T = 0.9$ with a MAP of 0.2710 (Fig. 1) and $\beta \geq 50$ with a MAP of 0.2762 (Fig. 2).

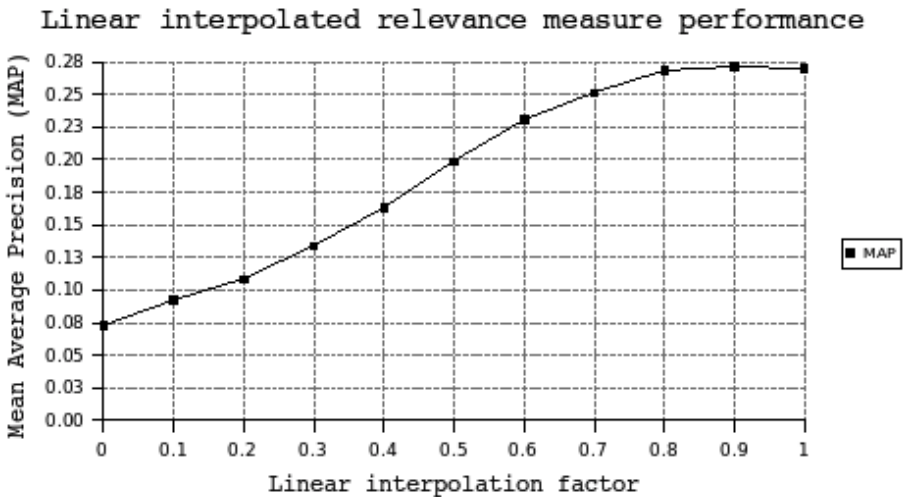


Fig. 1. Variation of Mean Average Precision (MAP) as a factor of linear interpolation factor λ_T (see Eq. 4)

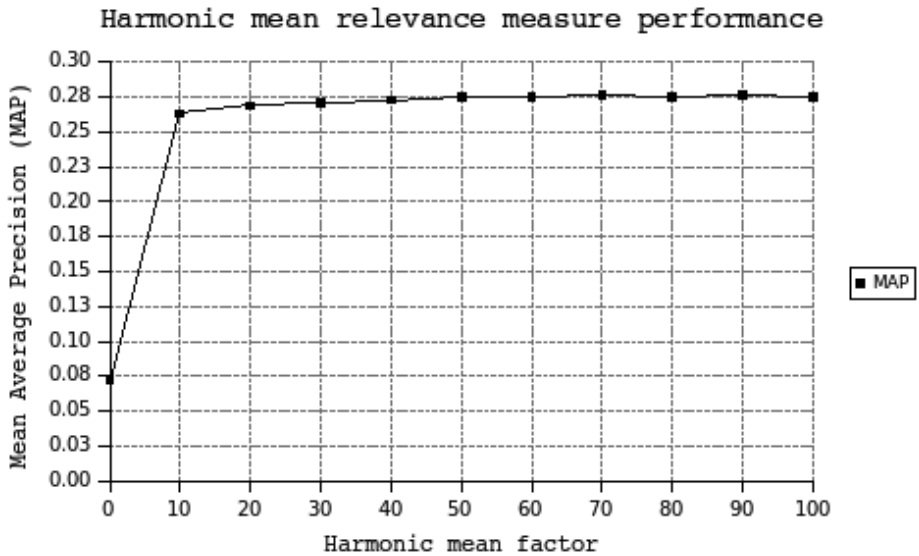


Fig. 2. Variation of Mean Average Precision (MAP) as a factor of harmonic mean factor β (see Eq. 6)

The harmonic mean based measure performs slightly better than the linear interpolation based measure. The best performance is achieved when the importance of non-geographic relevance outweighs the importance of geographic relevance. [8] reported an improvement when geographic terms in the query are weighed half or less than the weight of non-geographic terms.

Extended Harmonic Mean Combination. Figure 3 and Figure 4 show the performances of scope-based and type-based measures in comparison to the default Lucene performance using Eq.7. The scope-based and type-based relevance measures perform better than standard Lucene. The overall system performance (i.e. MAP) is 0.2941 compared to standard Lucene of 0.2695 which is a 9.1% improvement.

We can conclude that:

1. the performance of an IR system improves when geographic relevance is given a lesser importance.
2. geographic scope and type information can be used to construct a relevance measure to rank documents by geography giving better performance.
3. weighted harmonic mean combination of non-geographic and geographic relevance is a better approach than linear interpolated combination.

Possible factors affecting the performance include: (1) error in geo-tagging phase which feeds geo-scope analyser, and (2) error in geo-scope analysis phase. As future work we will address these areas, and improve on ranking formulas.

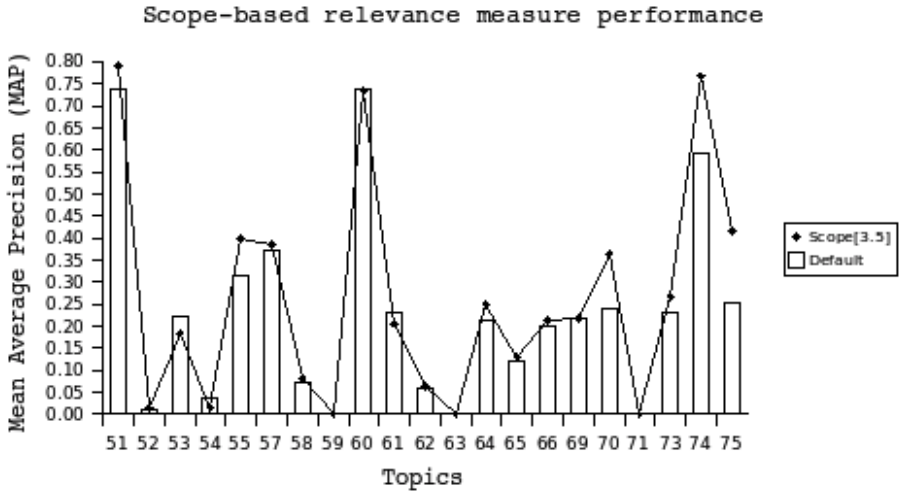


Fig. 3. Comparing scoped-based relevance measure performance (i.e. Eq. 7 with $GeoSim(q, d)$ replaced by $ScopeSim(q, d)$ with $\beta = 3.5$) against default Lucene performance (i.e. Eq. 3)

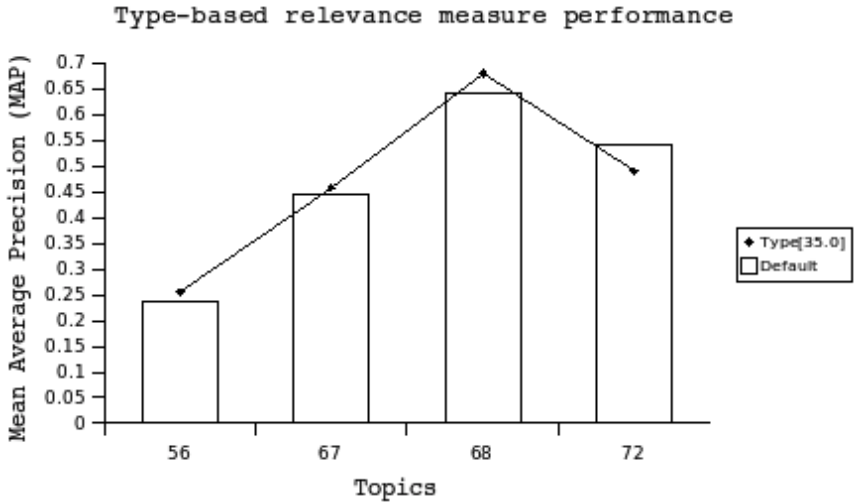


Fig. 4. Comparing type-based relevance measure performance (i.e. Eq. 7 with $GeoSim(q, d)$ replaced by $TypeSim(q, d)$ with $\beta = 35.0$) against default Lucene performance (i.e. Eq. 3)

4 Concluding Remarks

This paper defined scope-based and type-based relevance measures to rank documents by geographic restriction expressed in the user query. Both linear

interpolated and weighted harmonic mean combinations are used to combine non-geographic and geographic relevance scores. It is noted that harmonic mean combination slightly out performs the linear interpolated combination. It is further observed that to achieve better performance, the importance of non-geographic relevance need to outweigh the importance of geographic relevance.

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. Gospodnetic, O., Hatcher, E.: Lucene in Action. Manning Publications Co., Greenwich (2005)
2. Andrade, L., Silva, M.J.: Relevance Ranking for Geographic IR. In: Workshop on Geographical Information Retrieval, SIGIR 2006 (August 2006)
3. Andogah, G., Bouma, G., Nerbonne, J., Koster, E.: Resolving Geographical Scope of Documents with Lucene (unpublished, 2007)
4. van Rijsbergen, C.J.: Information Retrieval, 2nd edn., Butterworths, vol. 7, pp. 112–140 (1979)
5. Andogah, G., Bouma, G.: University of Groningen at GeoCLEF 2007. In: Working Notes for CLEF 2007 Workshop, Budapest, Hungary (September 2007)
6. Mandl, T., Gey, F., Nunzio, G.D., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xie, X.: GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Working Notes for CLEF 2007 Workshop, Budapest, Hungary (September 2007)
7. Gey, F., Larson, R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C.: GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Working Notes for CLEF 2006 Workshop, Alicante, Spain (September 2006)
8. Buscaldi, D., Rosso, P.: The UPV at GeoCLEF 2007. In: Working Notes for CLEF 2007 Workshop, Budapest, Hungary (September 2007)

Using Geographic Signatures as Query and Document Scopes in Geographic IR

Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva

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

Abstract. This paper reports the participation of the University of Lisbon at the 2007 GeoCLEF task. We adopted a novel approach for GIR, focused on handling geographic features and feature types on both queries and documents, generating signatures with multiple geographic concepts as a scope of interest. We experimented new query expansion and text mining strategies, relevance feedback approaches and ranking metrics.

1 Introduction

This paper presents the participation of the XLDB Group from the University of Lisbon at the 2007 GeoCLEF task. We experimented with novel strategies for geographic query expansion, text mining, relevance feedback and ranking metrics in a renewed GIR system. The motivation for this work derived from the results obtained in last year's participation, which revealed limitations on our previous GIR model [1].

First, our former GIR models focused on capturing and handling geonames and associated features for geographic reasoning, but ignored other terms with important geographic connotation, such as spatial relationships (e.g. in, near, on the shores of) and feature types (e.g. cities, mountains, airports). These terms may play an important role on the definition of the geographic relevance criteria of queries, and on the recognition of geonames in documents. At least, in the GeoCLEF 2007 topics, 13 out of the 25 topics of the Portuguese subtask contained feature types on the topic's title. So, for GeoCLEF 2007 we rebuilt the query processing modules so that all geographic information present on a query is captured, giving special attention to feature types and spatial relationships, as guides for the geographic query expansion [2].

Second, we rely on text mining methods to capture and disambiguate geonames extracted from the text, so that geographic scopes can be inferred for each document [3]. These methods involve geoname grounding into geographic concepts included in a geographic ontology, and disambiguation of hard cases through reasoning based on surrounding geonames also extracted from the text [1,4].

In CLEF 2006, we used a graph-ranking algorithm to analyse the captured features and assign one single feature as the scope of each document [5]. However, this proved to be too restrictive in some cases (other partial geographic contexts of the document were ignored), and also too brittle (incorrectly assigned scopes often lead to poor results). For example, too generic scopes were assigned to documents with geonames that do not correspond to adjacent areas: a document describing a football match between Portugal

and Hungary would have the common ancestor node (Europe) as a very strong candidate for final scope.

We therefore introduced a more comprehensive way to represent query and document scopes, generating geographic signatures for each document (D_{Sig}) and query (Q_{Sig}). A geographic signature is a list of geographic concepts that characterize a document or a query, allowing them to have several geographic contexts. The D_{Sig} is generated for each document by a text mining module, while the Q_{Sig} is generated through a geographic query expansion module. As a consequence of this novel geographic signature focused approach, the geographic ranking step now has the burden of evaluating relevance considering queries and documents with multiple geographic concepts as their scope, which required the development of new combination metrics for computing geographic relevance. In contrast, the similarity metric used last year only compared the (single) geographic concept as the scope of a document against the (single) geographic concept as the scope of a query.

The rest of this paper is organised as follows: Section 2 depicts our assembled GIR system, and describes in detail each module. Section 3 presents our experiments and analyses the results, and Section 4 ends with some conclusions and discussion topics.

2 System Description

Figure 1 presents the architecture of the GIR system assembled for GeoCLEF 2007, which has been presented in [6]. The GeoCLEF topics are automatically parsed by QueOnde and converted into *<what, spatial relationship, where>* triplets. The QuerCol module performs term and geographic query expansion, producing query strings consisting of query terms and a query geographic signature (Q_{Sig}). CLEF documents are loaded into a repository, becoming available to all modules. Faísca is a text mining module specially crafted to extract and disambiguate geonames, generating geographic signatures for each document (D_{Sig}). Sidra5 is the index and ranking module that generates text indexes from the documents and geographic indexes from their geographic signatures. Sidra5 also receives the queries generated by QuerCol as input, and generates final GeoCLEF runs in the *trec_eval* format.

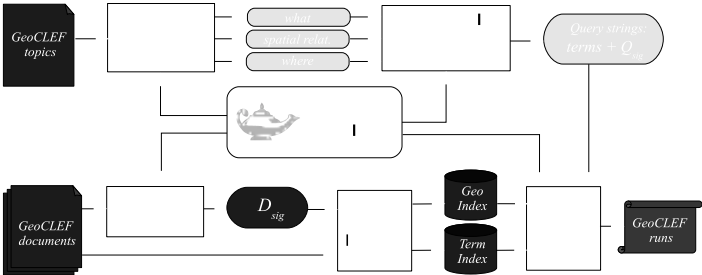


Fig. 1. Architecture of the GIR system assembled for GeoCLEF 2007

2.1 Geographic Ontology

All modules rely on a geographic ontology for geographic reasoning, created using our own geographic knowledge base, GKB [7]. The GKB 2.0 model now supports relationships between feature types, a better property assignment for features and feature types, and a better control of information sources [8]. Most of the ontology enrichment was carried out in the physical domain, with the addition of new feature types like airports, circuits and mountains, along with their instances.

For the purpose of our participation in GeoCLEF 2007, we made two improvements in the ontology: i) update of the GKB conceptual model to directly support multilingual names for geographic references, and ii) the addition of new features that we found missing after inspecting the GeoCLEF topics.

2.2 Query Parser and Query Expansion

We developed QueOnde, a new geographic query parsing module, which automatically converts query strings into *<what, spatial relationship, where>* triplets with the help of the geographic ontology and a set of manually-crafted context rules. These are used for capturing and disambiguating spatial relationships, features and feature types. For GeoCLEF, we consider the topic titles as query strings.

QuerCol is a geographic query expansion module, introduced in last year's participation [1,9]. QuerCol expands the thematic (*what*) and the geographic (*where*) parts of a query separately. The *what* is expanded through blind relevance feedback [10], while the *where* expansion is based on the available ontological information for the captured geographic concepts.

For GeoCLEF 2007, QuerCol was improved to handle feature types and spatial relationships, and to choose the appropriate geographic expansion strategy based on the features and feature types present in a query [2]. To better illustrate the reasoning task assigned to QuerCol, note that, when feature types are given in a query, they may mean two things: i) the user is disambiguating the geoname, because it can be associated to other geographic concepts (e.g., *City of Budapest* and *Budapest Airport*); or ii) the user is designating a set of concepts as a scope of interest (e.g., *Airports of Hungary*). In i), the feature type is disambiguating the geographic concept given by the feature *Budapest* as the scope of interest, while in ii), the feature type is designating a group of geographic concepts of the scopes of interest. QuerCol will choose the correct interpretation, and perform additional geographic reasoning to obtain the corresponding geographic concepts of the scope.

We now present a complete example of QueOnde and QuerCol integration to produce the Q_{Sig} : consider the following example taken from the GeoCLEF topic #74, *Ship traffic around Portuguese islands*: QuerOnde splits the topic title as a triplet, with *Ship traffic* as the thematic part, *in* as the spatial relationship, and *Portuguese islands* as the geographic part, sub-divided into *Portugal* as a grounded geoname and mapped into the corresponding ontological concept, and *islands* as a feature type. Given this query type, QueOnde therefore reasons that the scope of interest contains all geographic concepts of type *island* that have a *part-of* relationship with geographic concept *Portugal*. In the end, the Q_{Sig} is composed by the geographic concepts *São Miguel*, *Santa Maria*,

Formigas, Terceira, Graciosa, São Jorge, Pico, Faial, Flores, Corvo, Madeira, Porto Santo, Desertas and Selvagens.

2.3 Faísca

The text mining module Faísca parses the documents for geonames, generating the D_{Sig} . Faísca relies on a gazetteer of *text patterns* generated from the geographic ontology, containing all concepts represented by their feature name and respective feature types. The text patterns are in [*<feature type> \$ <feature name>*] and [*<feature name> <feature type>*] format (the former being more common in Portuguese texts, and the latter on English texts). Each pattern is assigned to a single *identifier* of the corresponding geographic concept in the ontology.¹ This immediately captures and grounds all geonames into their unique concept identifiers, without depending on hard-coded disambiguation rules. In the end, we have a *catch-all* pattern, which is used when the geoname found in the document does not contain any kind of external hints on its feature type. For these cases, we assign all identifiers of geographic concepts having that geoname.

The D_{Sig} generated by Faísca consists of a list of geographic concept identifiers and a corresponding *confidence measure* (*ConfMeas*) normalized to [0,1], representing the confidence on the feature being part of the document scope. *ConfMeas* is obtained through an analysis of the surrounding concepts on each case, in a similar way as described by Li et al. [11]. Geonames on a text are considered as qualifying expressions of a geographic concept when a direct ontology relationship between the geonames is also observed. For example, the geoname *Adelaide* receives an higher *ConfMeas* value on the document signature if an ontologically related concept, such as *Australia*, is nearby on the text. If so, the feature *Australia* is not included in the D_{Sig} , because it is assumed that it was used to disambiguate *Adelaide*, the more specific concept. Below is an example of D_{Sig} for document LA072694-001:

LA072694-0011: 5668[1.00]; 2230[0.33]; 4555[0.33]; 4556[0.33];

2.4 Sidra5

Sidra5 is a text indexing and ranking module with geographic capabilities based on MG4J [12]. It uses a standard inverted term index provided by MG4J, and a geographic forward index of [*docid*, D_{Sig}] that maps the id of a document to the corresponding D_{Sig} . Sidra5 first uses the *what* part of the query on the term index to retrieve the top 1000 documents. Afterwards, it retrieves the D_{Sig} of each document with the help of the geographic index. The document score is obtained by combining the Okapi BM25 *text score* [13], normalized to [0,1] (*NormBM25*) as defined by Song et al. [14], and a *geographic score* normalized to [0,1] (*GeoScore*) with equal weights:

$$Ranking(query, doc) = 0.5 \times NormBM25(query, doc) + 0.5 \times GeoScore(query, doc) \quad (1)$$

The calculation of *GeoScore* begins with the computation of the geographic similarity *GeoSim* for each pair (s_1, s_2), where s_1 in Q_{Sig} and s_2 in D_{Sig} , through a weighted sum of

¹ The character \$ means that an arbitrary term or group of terms is allowed to be present between the feature and the feature type, in order to avoid different stopword and adjective patterns.

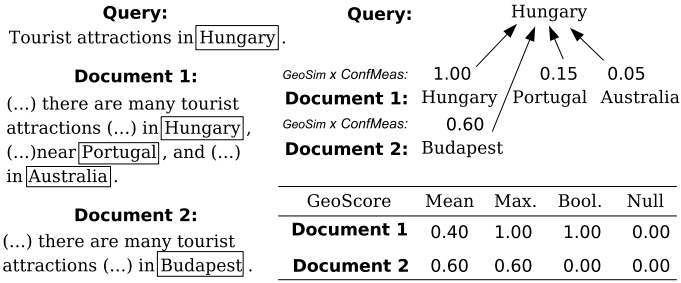


Fig. 2. Example of the computation of the four *GeoScore* combination metrics

four heuristic measures (discussed in our 2006 GeoCLEF participation [1]): Ontology (*OntSim*), Distance (*DistSim*), Adjacency (*AdjSim*) and Population (*PopSim*).

$$GeoSim(s_1, s_2) = 0.5 \times OntSim(s_1, s_2) + 0.2 \times DistSim(s_1, s_2) + 0.2 \times PopSim(s_1, s_2) + 0.1 \times AdjSim(s_1, s_2) \quad (2)$$

Having geographic signatures with multiple geographic concepts requires using aggregation metrics to calculate *GeoScore* from the different *GeoSim* values that a (*query*, *doc*) pair can generate. We experimented four metrics: Maximum, Mean, Boolean and Null.

GeoScore_{Maximum} is the maximum *GeoSim* value computed for a (*query*, *doc*) pair.

$$GeoScore_{Maximum}(query, doc) = \max(GeoSim(s_1, s_2) \times ConfMeas(s_2)), s_1 \in Q_{sig} \wedge s_2 \in D_{sig}$$

GeoScore_{Mean} is the average *GeoSim* values computed for a (*query*, *doc*) pair.

$$GeoScore_{Mean}(query, doc) = \text{avg}(GeoSim(s_1, s_2) \times ConfMeas(s_2)), s_1 \in Q_{sig} \wedge s_2 \in D_{sig}$$

GeoScore_{Boolean} equals 1 if there is a common concept in a (*query*, *doc*) pair, and equals 0 otherwise.

GeoScore_{Null} is always 0, turning off the geographic scores. This is used as a baseline metric for comparing results obtained with the other metrics.

The computation of the four *GeoScore* metrics is illustrated in Figure 2, which presents a fictional query and two document surrogates, along with the *GeoSim* × *ConfMeas* values and final *GeoScore* values.

3 Experiments and Results

Our experiments aimed at:

1. evaluating if this novel approach obtains better results than treating geonames as terms in a standard IR approach;
2. determining which *GeoScore* combination metrics is best.

3. measuring the importance of the geographic query expansion before or after the relevance feedback step.

All runs were generated in the following way: first, the topic titles are used for an initial retrieval, generating *initial runs*. The results of the initial runs are then used for query expansion through blind relevance feedback, generating final queries. These final queries are then used for a final retrieval, generating the *final runs*. More details on the run generation setup can be found in [6]. The generated runs represent three main experiments:

1. The *Terms only* experiment, that uses the names of the Q_{Sig} geographic concepts as standard terms in the generation of the initial and final runs. This means that this experiment uses only classical text retrieval. Nonetheless, the Q_{Sig} were generated by QuerCol through geographic query expansion.
2. The *Geo.QE* experiments, that uses text and geographic scores as described in Section 2.4. This experiment has two types of runs: *Geo. QE before RF*, where the fully expanded Q_{Sig} is used for the generation of the initial run and final run, and the *Geo. QE after RF*, that uses only the fully expanded Q_{Sig} on the generation of the final run; the initial run uses only the geographic concepts found on the initial query as the Q_{Sig} for the generation of the initial run.
3. The *Terms/GIR* experiment, that uses the initial run generated by the *Terms only* experiment to base the relevance feedback step, and afterwards uses the fully expanded Q_{Sig} for the generation of the final run, in the same way as the *Geo.QE* experiments generate their final runs after the relevance feedback step.

The results of our experiments are described on Table 1. We obtained significantly better results for the initial run by using geonames as terms instead of the respective geographic concepts (0.210 versus 0.126), which shows that this is an important result for the final results. The fact that the initial and final run of the *Terms Only* experiment was consistently better than the *Geo.QE* experiments, suggesting us to bootstrap a *Geo.QE* experiment with the initial run from the *Terms Only*, producing the *Terms/GIR* experiments. In the end, it obtained the highest MAP value from all our experiments (0.268 for the $GeoScore_{Boolean}$ metric).

Regarding the combination metrics, the $GeoScore_{Mean}$ produces poor MAP values because long document signatures tend to cause query drifting. $GeoScore_{Maximum}$ and $GeoScore_{Boolean}$ revealed to be much more robust, and the $GeoScore_{Boolean}$ metric has the best MAP values for Portuguese. This is explained in part because the $GeoScore_{Maximum}$ is highly dependent on the heuristics used, and these are dependent on the quality of the ontology, while the $GeoScore_{Boolean}$ metric is more straightforward on assigning maximum scores for geographically relevant documents.

We also noticed that using fully expanded Q_{Sig} produces better initial runs (0.126 versus 0.084 for Portuguese). This shows that the query signatures produced by QuerCol contribute to more relevant documents on the top of the retrieval results, which is helpful for the blind relevance feedback step. Yet, we did not observe this on the English subtask, prompting us to do further analysis to understand the reasons for this observation.

Table 1. MAP results obtained for the experiments

GeoScore		Terms only	Geo.QE before RF	Geo.QE after RF	Terms/GIR
Initial run		0.210	0.126	0.084	0.210
Final Run	Maximum	0.233	0.122	0.104	0.205
	Mean		0.022	0.021	0.048
	Boolean		0.135	0.125	0.268
	Null		0.115	0.093	0.221

a) Results for the Portuguese monolingual subtask.

Initial run		0.175	0.086	0.089	0.175
Final Run	Maximum	0.166	0.093	0.104	0.218
	Mean		0.043	0.044	0.044
	Boolean		0.131	0.135	0.204
	Null		0.081	0.087	0.208

b) Results for the English monolingual subtask.

4 Conclusions and Discussion

We tested a novel approach for GIR and evaluated its merits against standard IR approaches. We finally outperformed the standard IR approach, albeit in an unexpected way: the best experiment setup is to generate an initial run with classic text retrieval, and use the full geographic ranking modules for the generation of the final run. These results show that there are more efficient ways to introduce geographic reasoning on an IR system, and shed some light on what may be the main problem of many GIR approaches that fail to outperform standard IR approaches.

One should question if the full segregation of the thematic part and the geographic part, from query processing to document ranking, is really the best approach. In fact, as far as we know, there is no published work about a thorough evaluation on the effect of such segregation, claiming that this procedure clearly benefits GIR. A more detailed analysis showed that some terms added by relevance feedback were in fact geonames, and we noticed that geonames may also be good terms for standard IR.

An analysis for each topic reveals that our GIR system is very dependent on the quality of the geographic ontology, and has some limitations in the text mining step. For instance, 25% of all relevant documents (and, as such, with enough geographic evidence to define its scope) had an empty D_{Sig} . Also, we found that most geographic concepts found on the retrieved documents were not relevant for the document scope, or were not in the context of the topic. We also evaluated the results by query type, as the geographic query expansion shifted its strategy according to the spatial relationships, features and feature types found on the queries. We did not observe significative differences on the MAP values by query type.

As future work, we should revise our QR approach and use all query terms for the thematic and geographic expansion steps. The text mining module should also be improved to recognize more geonames and other named entities with a strong geographic connotation (e.g., monuments), and to better detect the roles of each geoname and its contribution for the scope of the document. In conclusion, a longer D_{Sig} does not imply a better D_{Sig} .

Acknowledgements

We thank Joana Campos for improving Faísca for the GIR prototype, Catarina Rodrigues for managing the geographic data, and Diana Santos for relevant suggestions. This work was jointly funded by the European Union (FEDER and FSE) and the Portuguese government, under contracts POSI/ISFL/13/408 (FIRMS-FCT) and POSC/339/1.3/C/NAC (Linguateca), and supported by grants SFRH/BD/29817/2006, POSI/SRI/47071/2002 (GREASE) and PTDC/EIA/73614/2006 (GREASE II) from FCT, 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., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 986–994. Springer, Heidelberg (2007)
2. Cardoso, N., Silva, M.J.: Query Expansion through Geographical Feature Types. In: 4th Workshop on Geographic Information Retrieval (GIR 2007), Lisbon, Portugal. ACM, New York (2007)
3. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P., Cardoso, N.: Adding Geographic Scopes to Web Resources. CEUS - Computers, Environment and Urban Systems 30, 378–399 (2006)
4. Cardoso, N., Martins, B., Andrade, L., Chaves, M.S., Silva, M.J.: The XLDB Group at GeoCLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., Müller, H., de Rijke, M. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 997–1006. Springer, Heidelberg (2006)
5. Martins, B., Silva, M.J.: A Graph-Based Ranking Algorithm for Geo-referencing Documents. In: Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), Houston, Texas, USA (2005)
6. Cardoso, N., Cruz, D., Chaves, M., Silva, M.J.: The University of Lisbon at GeoCLEF 2007. In: Peters, C., et al. (eds.) Working Notes of CLEF 2007, Budapest, Hungary (2007)
7. Chaves, M.S., Silva, M.J., Martins, B.: A Geographic Knowledge Base for Semantic Web Applications. In: Heuser, C.A. (ed.) Proceedings of the 20th Brazilian Symposium on Databases, Minas Gerais, Brazil, pp. 40–54 (2005)
8. Chaves, M.S., Rodrigues, C., Silva, M.J.: Data Model for Geographic Ontologies Generation. In: Ramalho, J.C., Lopes, J.C., Carriço, L. (eds.) XML: Aplicações e Tecnologias Associadas (XATA 2007), Lisbon, Portugal, pp. 47–58 (2007)
9. Cardoso, N., Silva, M.J., Martins, B.: The University of Lisbon at CLEF 2006 Ad-Hoc Task. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 51–56. Springer, Heidelberg (2007)
10. 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)
11. Li, Y., Moffat, A., Stokes, N., Cavedon, L.: Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In: 3rd Workshop on Geographical Information Retrieval (GIR 2006), Seattle, Washington, USA (2006)

12. Boldi, P., Vigna, S.: MG4J at TREC 2005. In: Proceedings of the 14th Text REtrieval Conference (TREC 2005), NIST SP 500-266 (2005), <http://mg4j.dsi.unimi.it>
13. Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A., Lau, M.: Okapi at TREC-3. In: Proceedings of the 3rd Text REtrieval Conference (TREC-3), pp. 21–30 (1992)
14. Song, R., Ji-RongWen, S.S., Xin, G., Tie-YanLiu, Q.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 2004) (2004)

Cheshire at GeoCLEF 2007: Retesting Text Retrieval Baselines

Ray R. Larson

School of Information, University of California, Berkeley, USA
`ray@ischool.berkeley.edu`

Abstract. In this paper we will briefly describe the approaches taken by Berkeley for the main GeoCLEF 2007 tasks (Mono and Bilingual retrieval). The approach this year was to use probabilistic text retrieval based on logistic regression and incorporating blind relevance feedback for all of the runs. Our intent was to establish a baseline result without explicit geographic processing for comparison with future geographic processing approaches. All translation for bilingual tasks was performed using the LEC Power Translator machine translation system.

1 Introduction

This paper very briefly describes the retrieval algorithms and evaluation results for Berkeley's official submissions for the GeoCLEF 2007 track. All of the runs were automatic without manual intervention in the queries (or translations). We submitted three Monolingual runs (one German, one English, and one Portuguese) and nine Bilingual runs (each of the three main languages to each other language, and three runs from Spanish to English, German, and Portuguese).

This paper first describes a key aspect of 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 analysis and comparison with previous years, and finally present conclusions and future directions for GeoCLEF participation.

2 The Retrieval Algorithms

The basic form and variables for the *Logistic Regression* (LR) algorithm used for all of our submissions (for GeoCLEF, ImageCLEFPhoto, and for Domain Specific is described in detail in our full-length paper for the Domain Specific track in this volume. Here we describe only the blind relevance feedback approach that was also used in retrieval for all three of the tracks.

The algorithm used for blind feedback was originally developed and described by Chen [1]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [3].

Table 1. Contingency table for term relevance weighting

	Relevant	Not Relevant	
In doc	R_t	$N_t - R_t$	N_t
Not in doc	$R - R_t$	$N - N_t - R + R_t$	$N - N_t$
	R	$N - R$	N

Blind relevance feedback is 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. Obviously there are important choices to be made regarding the number of top-ranked documents to consider, and the number of terms to extract from those documents. For GeoCLEF we chose to use the top ten terms from ten top-ranked documents. The terms were chosen by extracting the document vectors for each of the ten and computing the Robertson and Sparck Jones term relevance weight for each document. This weight is based on a contingency table where the counts of four different conditions for combinations of (assumed) relevance and whether or not the term is, or is not in a document. Table 1 shows this contingency table.

The relevance weight is calculated using the assumption that the first ten documents are relevant and all others are not. For each term in these documents the following weight is calculated:

$$w_t = \log \frac{\frac{R_t}{R - R_t}}{\frac{N_t - R_t}{N - N_t - R + R_t}} \quad (1)$$

The ten terms (including those that appeared in the original query) with the highest w_t are selected and added to the original query terms. For the terms not in the original query, the new “term frequency” (qtf_i in main LR equation shown in our Domain Specific paper in this volume) is set to 0.5. Terms that were in the original query, but are not in the top ten terms are left with their original qtf_i . For terms in the top ten and in the original query the new qtf_i is set to 1.5 times the original qtf_i for the query. The new query is then processed using the same Logistic Regression algorithm and the ranked results returned as the response for that topic.

3 Approaches for GeoCLEF

Although the Cheshire system permits geographically controlled indexing of texts (i.e., all proper nouns are looked up a gazetteer and only those with matches in the gazetteer are indexed, or have the appropriate coordinates inserted in the index) these indexes were not used in the submissions this year.

Table 2. Submitted GeoCLEF Runs: runs that had the highest overall MAP for the task have asterisks next to the run name

Run Name	Description	Type	MAP
BerkMODEBASE	Monolingual German	TD auto	0.1392
BerkMOENBASE*	Monolingual English	TD auto	0.2642
BerkMOPTBASE	Monolingual Portuguese	TD auto	0.1739
BerkBIENDEBASE	Bilingual English⇒German	TD auto	0.0902
BerkBIENPTBASE	Bilingual English⇒Portuguese	TD auto	0.2012
BerkBIDEENBASE*	Bilingual German⇒English	TD auto	0.2208
BerkBIDEPTBASE	Bilingual German⇒Portuguese	TD auto	0.0915
BerkBIPTDEBASE	Bilingual Portuguese⇒German	TD auto	0.1109
BerkBIPTENBASE	Bilingual Portuguese⇒English	TD auto	0.2112
BerkBIESDEBASE	Bilingual Spanish⇒German	TD auto	0.0724
BerkBIESENBASE	Bilingual Spanish⇒English	TD auto	0.2195
BerkBIESPTBASE	Bilingual Spanish⇒Portuguese	TD auto	0.1924

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. However, the German language runs *did not* use compounding in the indexing and querying processes to generate simple word forms from compounds. Although we tried again this year to make this work within the Cheshire system, we again lacked the time needed to implement it correctly.

The Snowball stemmer was used by Cheshire for language-specific stemming.

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 (English, German, and Portuguese), for bilingual tasks the topics were translated from the source language to the target language using the LEC Power Translator PC-based machine translation system[2]. In all cases the “title” and “desc” topic elements were combined into a single probabilistic query. We consider all of these runs to be the simplest “baseline” for our system, using no combinations of indexes or other complex processing and relying only on the LR algorithm with blind feedback operating on the entire text contents of the collections.

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 2.

Once again we found some rather anomalous results among the official runs. For example, it is not at all clear, given the same basic approach used for all of the runs, why the bilingual runs for English⇒Portuguese (MAP 0.2012), and Spanish⇒Portuguese (MAP 0.1924) should have performed better than our Monolingual Portuguese run (MAP 0.1739).

Table 3. Comparison of Berkeley's best 2005 and 2006 runs for English and German

TASK	MAP 2006	MAP 2007	Pct. Diff.
Monolingual English	0.2499	0.2642	5.7222
Monolingual German	0.2151	0.1392	-54.5259
Monolingual Portuguese	0.1622	0.1739	7.2133
Bilingual English⇒German	0.1561	0.0902	-73.0599
Bilingual English⇒Portuguese	0.12603	0.2012	59.6825

Obviously the “weak man” in our current implementation is German. This may be due to compounding issues, but the lower results are clear in both Monolingual and Bilingual runs where either the source topics or the target data is German.

5 Analysis and Conclusions

Because we used a virtually identical processing approach (except for translation) this year as we used for some of our runs submitted for GeoCLEF 2006, we build Table 3 examine the differences. Overall, we did see some improvements in results. However, the submitted 2006 results used compounding for German, which would appear to be the primary cause of our declining monolingual and bilingual scores for German, although the translation software may also be at fault. Otherwise, our bilingual results this year are largely due to the effectiveness of our new translation software. We used the Spanish topic statements provided for bilingual Spanish to English, German, and Portuguese, and saw results that look quite good for English and Portuguese, exception again being German.

Although we did not do any explicit geographic processing for this year, we plan to do so again 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. In addition, we used only the title and desc elements of topics this year, and also we did not use automatic expansion of toponyms in the topic texts. Since this was done explicitly in some of the topic narratives we may have missed possible improvements by not using the entire topic. In previous years it has been apparent that implicit or explicit toponym inclusion in queries, as might be expected, leads to better performance when compared to using titles and descriptions alone in retrieval.

References

1. 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)
2. LEC Power Translator 11 Premium, <http://www.lec.com>
3. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science, 129–146 (May–June, 1976)

On the Relative Importance of Toponyms in GeoCLEF

Davide Buscaldi and Paolo Rosso

Natural Language Engineering Lab (NLE Lab)
Dpto. de Sistemas Informáticos y Computación (DSIC)
Universidad Politécnica de Valencia, Spain
{dbuscaldi,proso}@dsic.upv.es

Abstract. In this work we attempted to determine the relative importance of the geographical and WordNet-extracted terms with respect to the remainder of the query. In our system, geographical terms are expanded with WordNet holonyms and synonyms and indexed separately. We checked the relative importance of the terms by multiplying their weight by 0.75, 0.5 and 0.25. The comparison to the baseline system, which uses only Lucene, shows that in some cases it is possible to improve the mean average precision by balancing the relative importance of geographical terms with respect to the content words in the query. We also observed that WordNet holonyms may help in improving the recall but WordNet has a small coverage and term expansion is sensible to ambiguous place names.

1 Introduction

Since our first participation at the GeoCLEF we have been developing a method that can use the information contained in the WordNet [1] ontology for the Geographical Information Retrieval task. In our first attempt [2,3] we simply used synonyms (alternate names) and meronyms of locations that appeared in the query in order to expand the query itself. This method performed poorly, due to the noise introduced by the expansion. Subsequently, we introduced a method that exploits the inverse of the meronymy relationship - *holonymy* (a concept *A* is *holonym* of another concept *B* if *A* contains *B*). We named this method *Index Term Expansion* [4]. With this method we add to the geographical index terms the information about their holonyms, such that a user looking for information about *Spain* will find documents containing *Valencia*, *Madrid* or *Barcelona* even if the document itself does not contain any reference to Spain. The results obtained with this method showed that the inclusion of WordNet holonyms allowed to obtain an improvement in recall. Moreover, we noticed that the use of the Index Term Expansion method did not allow to obtain the same precision of the baseline system. We individuated the reason of this behaviour in the fact that the geographical terms were assigned the same importance as the other terms of the query. Therefore, in this participation we attempted to determine the relative importance of geographical and WordNet-extracted terms

with respect to the remainder of the terms of the query. This has been done by means of the separation of the index of geographical terms from the general index and the creation of another index that contains only WordNet-extracted terms.

Another possible reason for the bad performance of Index Term Expansion is the ambiguity of toponyms. This is a common problem in news text [5], and currently various approaches are being developed [6,7]. We attempted to determine how many of the toponyms appearing in topics were ambiguous, according to WordNet.

In the following section, we describe the system and how index term expansion works. In section 3 we describe the characteristics of our submissions and show a resume of the obtained results. Finally, we draw some conclusions and discuss further work.

2 Our GIR System

The core of the system is constituted by the Lucene¹ open source search engine, version 2.1. The engine is supported by a module that uses LingPipe² for HMM-based Named Entity recognition (this module performs the task of recognizing geographical names in texts), and another one that is based on the MIT Java WordNet Interface³ in order to access the WordNet ontology and find synonyms and holonyms of the geographical names.

2.1 Indexing

During the indexing phase, the documents are examined in order to find location names (*toponyms*) by means of LingPipe. When a toponym is found, then two actions are performed: first of all, the toponym is added to a separate index (*geo* index) that contains only the toponyms. The separation of indices is commonly used in GIR and it has been demonstrated to improve the retrieval results [8,9]. We introduced a third index that contains only WordNet-related information. In the next step WordNet is examined in order to find holonyms (recursively) and synonyms of the toponym. The retrieved holonyms and synonyms are put in this separate index (*wn* index).

For instance, consider the following text from the document GH950630-000000 in the Glasgow Herald 95 collection:

...The British captain may be seen only once more here, at next month's world championship trials in Birmingham, where all athletes must compete to win selection for Gothenburg...

The following toponyms are added to the *geo* index: "Birmingham", "Gothenburg". Birmingham is found in WordNet both as *Birmingham*, *Pittsburgh of the*

¹ <http://lucene.apache.org/>

² <http://www.alias-i.com/lingpipe/>

³ <http://www.mit.edu/~markaf/projects/wordnet/>

South, in the United States and *Birmingham*, *Brummagem*, an important city in England. The holonyms in the first case are *Alabama*, *Gulf States*, *South*, *United States of America* and their synonyms. In the second case, we obtain *England*, *United Kingdom*, *Europe* and their synonyms. All these words are added to the *wn* index for Birmingham, since we did not use any method in order to disambiguate the toponym. For Gothenburg we obtain *Sweden* and *Europe* again, together with the original Swedish name of Gothenburg (*Goteborg*). These words are also added to the *wn* index.

In Figure 1 we show the terms that are assigned to each of the indices in Lucene for the above text.

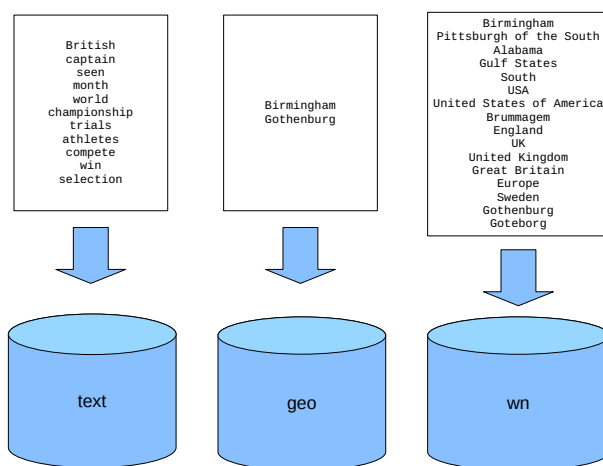


Fig. 1. Repartition of terms of the example text in the indices

2.2 Searching

For each topic, LingPipe is run again in order to find the geographical terms. In the search phase, we do not use WordNet. However, the toponyms individuated by LingPipe are searched in the geographical and/or WordNet indices. All query terms are separated by a standard OR operator.

3 Experiments

We submitted a total of 12 runs at GeoCLEF 2007. Two runs were used as benchmarks: they were obtained by using the baseline system, without index term expansion, in one case considering only topic title and description, and all fields in the other case. The remaining runs used the *geo* index or *wn* index or both, with different weightings that were submitted using the Lucene “Boost” operator. This operator allows to assign relative importance to terms. This means that a term with, for instance, a boost factor of 4 will be four times more

important than the other terms in the query. We used 0.75, 0.5 and 0.25 as boost factor for geographical and WordNet terms, in order to study their importance in the retrieval process.

In the following tables we show the results obtained in terms of Mean Average Precision and Recall for all the submitted runs.

Table 1. Mean Average Precision (MAP) and Recall obtained for all the “Title+Description only” runs

run ID	geo boost	wn boost	MAP	Recall
rfiaUPV01	0	0	0.226	0.886
rfiaUPV03	0.5	0.0	0.227	0.869
rfiaUPV05	0.5	0.25	0.238	0.881
rfiaUPV07	0.75	0.0	0.224	0.860
rfiaUPV08	0.75	0.25	0.224	0.860
rfiaUPV09	0.25	0.25	0.239	0.888
rfiaUPV10	0.25	0.0	0.236	0.891
rfiaUPV11	0.5	0.5	0.239	0.886
rfiaUPV12	0.75	0.75	0.231	0.877

Table 2. Mean Average Precision and Recall obtained for the “All fields” runs

run ID	geo boost	wn boost	MAP	Recall
rfiaUPV02	0	0	0.247	0.903
rfiaUPV04	0.5	0.0	0.256	0.915
rfiaUPV06	0.5	0.25	0.263	0.926

The results obtained with the topic title and description (Table 1) show that by reducing the weight of geographical terms with respect to content words it is possible to obtain a better MAP. The integration of WordNet terms allows to improve further the MAP, although it has almost no effect over recall. However, if we consider all the topic fields (Table 2) we can observe that the introduction of WordNet allowed to improve also the recall.

We carried out a topic-by-topic analysis in order to understand a reason for this puzzling behaviour. We discovered that our method allowed to improve the results especially for the following topics: 51, 64, 74 and 75. The baseline system performed better on topics 62 and 66. Figure 2 shows these topics in detail. The Mean Average Precision (MAP) obtained for these topics with our system are shown in Table 3.

In topic 62 the long list of countries in the narrative unbalanced the search towards “Eastern Europe countries”, reducing the importance of “OSCE meetings”. This is due mostly to the weighting scheme adopted ($tf \cdot idf$). In topic 66 the problem is that *Bosphorus* is not present in WordNet and therefore it could not be expanded.

```

<num>10.2452/51-GC</num>
<title>Oil and gas extraction found between the UK and the
Continent</title>
<desc>To be relevant documents describing oil or gas production
between the UK and the European continent will be relevant</desc>
<narr>Oil and gas fields in the North Sea will be relevant.</narr>

<num>10.2452/62-GC</num>
<title>OSCE meetings in Eastern Europe</title>
<desc>Find documents in which Eastern European conference
venues of the Organization for Security and Co-operation in
Europe (OSCE) are mentioned</desc>
<narr>Relevant documents report on OSCE meetings in Eastern Europe.
Eastern Europe includes Bulgaria, Poland, the Czech Republic,
Slovakia, Hungary, Romania, Ukraine, Belarus, Lithuania, Estonia,
Latvia and the European part of Russia.</narr>

<num>10.2452/64-GC</num>
<title>Sport events in the french speaking part of
Switzerland</title>
<desc>Find documents on sport events in the french speaking part
of Switzerland</desc>
<narr>Relevant documents report sport events in the french speaking
part of Switzerland. Events in cities like Lausanne, Geneva,
Neuchtel and Fribourg are relevant.</narr>

<num>10.2452/66-GC</num>
<title>Economy at the Bosphorus</title>
<desc>Documents on economic trends at the Bosphorus strait</desc>
<narr>Relevant documents report on economic trends and development
in the Bosphorus region close to Istanbul</narr>

<num>10.2452/74-GC</num>
<title>Ship traffic around the Portuguese islands</title>
<desc>Documents should mention ships or sea traffic connecting
Madeira and the Azores to other places, and also connecting the
several isles of each archipelago. All subjects, from wrecked ships,
treasure finding, fishing, touristic tours to military actions, are
relevant, except for historical narratives.</desc>
<narr>Documents have to mention that there is ship traffic connecting
the isles to the continent (portuguese mainland), or between the
several islands, or showing international traffic. Isles of Azores
are: Sao Miguel, Santa Maria, Formigas, Terceira, Graciosa, Sao
Jorge, Pico, Faial, Flores and Corvo. The Madeira islands are:
Mardeira, Porto Santo, Desertas islets and Selvagens islets.</narr>

<num>10.2452/75-GC</num>
<title>Violation of human rights in Burma</title>
<desc>Documents are relevant if they mention actual violation of
human rights in Myanmar, previously named Burma.</desc>
<narr>This includes all reported violations of human rights in Burma,
no matter when (not only by the present government).
Declarations (accusations or denials) about the matter only,
are not relevant.</narr>

```

Fig. 2. The analysed topics

Table 3. Comparison of the MAP obtained for the topics in Fig. 2 (All-fields runs)

topic	Baseline	RFIAUPV04 (geo)	RFIAUPV06 (geo+WN)
51	67.73	71.86	77.59
64	10.55	23.05	18.11
74	44.38	60.90	57.57
75	39.72	47.52	55.11
62	12.15	4.70	6.08
66	37.92	24.67	23.52

The comparison of the results of the runs based on WordNet and those that do not consider WordNet showed that topics 74 and 64 were those that did not take advantage from the holonym expansion. In the first case it was due to the absence of meronyms for the two major Portuguese archipelagos, the Azores and Madeira. In the second one, it was due to the presence of meronyms for Switzerland that are not related to the French-speaking part of the country.

The relative importance of geographical terms depends on the topic contents: in Table 4 we can observe that increasing the weight of geographical terms was effective especially for topic 74, while it greatly reduced MAP for topic 62. It can be corrected either by recurring to a different weighting scheme, or by performing an accurate topic analysis.

Table 4. Comparison of the MAP obtained for the topics in Fig. 2 at different weights for geographical terms (Topic+Description runs)

topic	Baseline	<i>geo</i> * 0.25	<i>geo</i> * 0.5	<i>geo</i> * 0.75
51	52.90	52.10	51.17	48.95
64	1.10	2.08	2.38	2.42
74	37.63	47.88	68.18	72.73
75	36.38	48.66	46.57	46.12
62	35.30	33.78	6.07	2.70
66	36.68	36.68	36.68	36.68

In order to check the impact of ambiguity over the results, we counted how many toponyms in the topics were present in WordNet and how many of them were ambiguous: we found that of 93 toponyms, only 59 were present in WordNet (corresponding to a coverage of 63.4%), with 87 different senses (average polysemy: 1.47 senses per toponym).

4 Conclusions and Further Work

The obtained results show that it is necessary to find a balance between the geographical terms and content words in the topics, and this balance often depends

on the topic itself. Reducing the importance of geographical terms allowed to improve the mean average precision in some cases. The use of WordNet may be helpful for some topics, even if in other cases it does not help or introduces errors. The coverage of WordNet is rather small with respect to the toponyms found in the queries; the use of a larger geographical resource (such as the Geonet Names Server⁴ gazetteer or the Getty thesaurus of Geographical Names⁵) may allow to improve the coverage. However, we observed that WordNet toponyms are ambiguous on average, and ambiguity will surely increase when using a more detailed resource. We plan to integrate a toponym disambiguation method in our next participation to GeoCLEF in order to overcome the ambiguity issue. We are going to implement also an effective query analysis method, in order to use different search strategies (and term weightings) depending on the input query.

Acknowledgements

We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work.

References

1. Miller, G.A.: Wordnet: A lexical database for english. *Communications of the ACM* 38, 39–41 (1995)
2. 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., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2005. LNCS*, vol. 4022, pp. 939–946. Springer, Heidelberg (2006)
3. Gey, F.C., 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, pp. 908–919. Springer, Heidelberg (2006)
4. Buscaldi, D., Rosso, P., Sanchis, E.: A WordNet-based Indexing Technique for Geographical Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) *CLEF 2006. LNCS*, vol. 4730, pp. 954–957. Springer, Heidelberg (2007)
5. Garbin, E., Mani, I.: Disambiguating toponyms in news. In: *Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005)*, Morristown, NJ, USA, pp. 363–370. Association for Computational Linguistics (2005)
6. Overell, S., Rüger, S.: Geographic co-occurrence as a tool for GIR. In: *GIR 2007: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pp. 71–76. ACM, New York (2007)

⁴ <http://earth-info.nga.mil/gns/html/index.html>

⁵ http://www.getty.edu/research/conducting_research/vocabularies/tgn/

7. Buscaldi, D., Rosso, P.: A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Systems* (accepted, to be published, 2008)
8. Leveling, J., Veiel, D.: Experiments on the Exclusion of Metonymic Location Names from GIR. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006. LNCS*, vol. 4730, pp. 901–904. Springer, Heidelberg (2007)
9. Cardoso, N., Silva, M.J.: Query expansion through geographical feature types. In: *GIR 2007: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pp. 55–60. ACM, New York (2007)

Filtering for Improving the Geographic Information Search

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega,
and L.A. Ureña-López

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

Abstract. This paper describes the GEOUJA System, a Geographical Information Retrieval (GIR) system submitted by the SINAI group of the University of Jaén in GeoCLEF 2007. The objective of our system is to filter the documents retrieved from an *information retrieval* (IR) subsystem, given a multilingual statement describing a spatial user need. The results of the experiments show that the new heuristics and rules applied in the *geo-relation validator* module improve the general precision of our system. The increasing of the number of documents retrieved by the *information retrieval* subsystem also improves the final results.

1 Introduction

This paper describes the second participation of the SINAI¹ research group of the University of Jaén in GeoCLEF 2007[1]. In GeoCLEF 2006 we studied the behavior of query expansion[2]. The results showed us that the expansion of topics did not improve the baseline case. However, the results obtained in GeoCLEF 2007 make clear that filtering documents retrieved increases the precision and the recall of a geographical information search.

In GeoCLEF 2007 the results obtained showed that the heuristics applied were quite restrictive. In the post experiments presented in this paper we have defined additional rules and heuristics, less restrictive than used in the official task. In the *geo-relation validator* module, the most important subsystem in our architecture, we have eliminated the heuristic that considered entities appearing in query without an associated *geo-relation*. In addition, the number of retrieved documents by the *information retrieval* subsystem has been increased too, in order to provide a larger variety of documents to be checked by the *geo-relation validator* subsystem.

The next section describes the system overview. Then, each module of the system is explained in the section 3. In the section 4, experiments and results are described. Finally, the conclusions and future work are expounded.

¹ <http://sinai.ujaen.es>

2 System Overview

We propose a Geographical Information Retrieval System made up of five related subsystems. These modules are explained in detail in the next section.

In our architecture we only worked with the English collections, *Los Angeles Times 94* (LA94) and *Glasgow Herald 95* (GH95). The collections have been pre-processed off-line (English *stop-words* list, a named entity recognizer (NER) and the Porter *stemmer*[3]). The pre-processed data set is indexed using the **information retrieval subsystem** (IR subsystem).

Bilingual topics are translated by means of the **translation subsystem**. Then, we labeled them with NER and *geo-relation* information. The **geo-relation finder subsystem** (GR finder subsystem) extracts spatial relations from the geographic query and the **NER subsystem** recognizes named entities.

The IR subsystem returns a list of relevant documents for the original English query. The NER subsystem extracts only the locations from this list. These locations per document, the entities and the geo-references founded in topics, are the input for the **geo-relation validator subsystem** (GR validator subsystem), the main module in our architecture.

The GR validator subsystem eliminates those documents previously retrieved that do not satisfy the predefined rules. These rules are related to all the information that this module handles (locations and spatial relations from documents and geographic queries) and are explained in section 3.4. Figure 1 shows the proposed system architecture.

3 Subsystems Description

3.1 Translation Subsystem

As translation module we have used SINTRAM (SINai TRANslation Module)[4]. This subsystem translates the queries from several languages into English. SINTRAM uses some online Machine Translators for each language pair and implements some heuristics to combine the different translations. A comprehensive evaluation showed that Systran² performed best for German and Portuguese.

3.2 Named Entity Recognizer Subsystem

The main goal of NER subsystem is to detect and recognize the entities appearing in the collections and the queries. We are only interested in geographical information, so we have just used *locations* detected by this NER module. We have worked with the NER module of the GATE³ toolkit. The location terms include everything that is town, city, capital, country and even continent. The NER module adds *entity labels* to the topics with the found locations. Each entity recognized in a topic is labeled as follows:

² <http://www.systransoft.com>

³ <http://gate.ac.uk/>

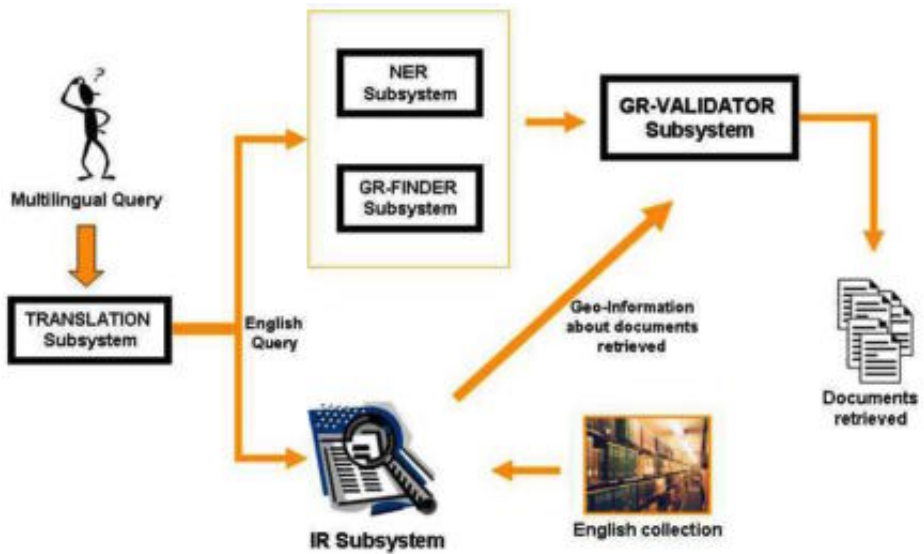


Fig. 1. The GEOUJA System architecture

```
<entity pos="15" type="LOC"> USA </entity>
```

where *pos* is the position of the entity in the phrase. This value is greater than or equal to zero and we used it to know at any moment what locations and *geo-relations* are related to each other by proximity. These topic labels will be used later by the GR finder subsystem.

3.3 Geo-relation Finder Subsystem

The *geo-relation* finder subsystem is used to find the spatial relations in the geographic queries. This module makes use of four text files to store the *geo-relations* identified. Four *geo-relations* files exist because our system detects spatial relations of four words at the most. Some *geo-relations* examples are: *in*, *near*, *north of*, *next to*, *in or around*, *in the west of*...

The GR finder module adds *geo-relation* labels to the topics. A *geo-relation* label example recognized in a topic would be the following one:

```
<gr pos="30" long="1" value="in">
  <entity pos="38"> Edinburgh </entity>
</gr>
```

where *pos* is the position of the spatial relation in the phrase, *long* is the number of words in the spatial relation and *value* is the *geo-relation*.

In this module we controlled a special *geo-relation* named *between*. For this case, the GR finder subsystem adds the two entities that this preposition relates. An example of the label this module adds for the text "To be relevant documents

describing oil or gas production between the UK and the European continent will be relevant” is:

```
<gr pos="9" long="1" value="between">
  <entity pos="11"> UK </entity>
  <entity pos="14"> European continent </entity>
</gr>
```

The basic operation of the GR finder subsystem is to found the spatial relation related with each entity recognized by NER subsystem. If an entity is not related with any *geo-relation*, the module will not include that entity in the text file that will be used later by the GR validator subsystem. Sometimes, an entity with its spatial relation are repeated in the same topic. This module will include only one time the entity with its spatial relation for improving the runtime of the GR validator subsystem later.

3.4 Geo-relation Validator Subsystem

This is the most important module of our system. Its main goal is to discriminate what documents among the recovered ones by the IR subsystem are valid. This module makes use of the location classification submodule.

Location Classification Submodule. This module is used by the GR validator subsystem. Its objective is to define the location type of the entity from the query, that will be used subsequently in the heuristics. The location type can be: *continent*, *country*, *city* or *place*. To make the classification, the module uses the Geonames gazetteer. First, it checks if the entity is a continent. If not, the module identifies whether the entity is a country, looking into a vector containing all the names of worldwide countries. If the entity is not a continent or a country, then the system gets the *featureClass* attribute from the Geonames gazetteer. If this attribute is equal to "P" or "A" indicates that the entity is a city. In another case, the entity is considered "place".

In order to apply different heuristics, the GR validator module makes use of geographical data. This *geo-information* has been obtained from the Geonames Gazetteer⁴. This module solves queries like:

- Find the country name of a city.
- Find the latitude and longitude for a given location.
- Check if a city belongs to a certain country.
- Check if a location is to the north of another one.

Many heuristics can be applied to make the validation of a document recovered by IR subsystem. The GR validator subsystem receives external information from the IR subsystem (entities from each document recovered) and from GR finder and NER subsystems (entities and spatial relations from each topic). We have used the following heuristics in our experiments:

⁴ <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.

- If the entity that appears in a topic is a **country** and it has associated "*in the north of*" as spatial relation, the module obtains from the Geonames Gazetteer the maximum and minimum latitudes of all locations that belong to that country. Then, the module calculates half of the latitude from the maximal and minimal latitudes to estimate the north of a region. Any location that is above the half latitude will be considered in the northern part of the country.
- If the entity that appears in a topic is a **city** and it has associated "*near to*" as spatial relation, we have considered that a location is *near to* another when it is at a distance of less than 50 kilometers around. To measure the distance in kilometers between two locations we have used the simple formula for calculating the distance with geographic coordinates:

$$d = \sqrt{(x^2 + y^2)}$$

where:

$$x = 110.56 * \text{abs}(\text{lat2} - \text{lat1})$$

$$y = 84.8 * \text{abs}(\text{lon2} - \text{lon1})$$

lat2, *lat1*, *lon2* and *lon1* are the latitudes and longitudes from location 2 and 1 respectively. If the *d* value is less than 50, the location 1 will be considered *near to* location 2.

- If the entity that appears in a topic is a **continent** or a **country** and it has associated "*in*", "*of*", "*at*", "*on*", "*from*" or "*along*" as spatial relation, the module will accept the document if a location exists in the document that belongs to that continent or country.

We only have used the heuristic for the north direction because only this case appears in the topics. New heuristics will be implemented in the future for other directions.

3.5 Information Retrieval Subsystem

The information retrieval system that we have employed is Lemur⁵. One parameter for each experiment has been the weighting function, such as Okapi[5] or *TF.IDF*. Another has been the inclusion or exclusion of Pseudo-Relevant Feedback (PRF)[6].

4 Experiments and Results

The SINAI group has participated in monolingual and bilingual tasks for Geo-CLEF 2007 with a total of 26 experiments. We have considered all tags from topics (*title*, *description* and *narrative*) for the information retrieval process.

⁵ <http://www.lemurproject.org/>. 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.

Our baseline experiment, without applying heuristics or rules on relevant documents retrieved, has been applied in the monolingual and bilingual tasks.

In post experiments, we have applied the additional heuristics and rules less restrictive from the GR validator subsystem (see section 3.4) and we have also explored a larger number of retrieved documents by the IR subsystem, in the aim of providing a larger variety of documents to be checked by the GR validator subsystem. Some results are shown in Table 1.

Table 1. Summary of results

Experiment	Docs. Retrieved	Weighting Function	PRF	MAP
Baseline (CLEF07)	1000	Okapi	no	0.2486
Baseline (CLEF07)	1000	TF.IDF	no	0.1777
Baseline (CLEF07)	1000	Okapi	yes	0.2605
Baseline (CLEF07)	1000	TF.IDF	yes	0.1803
Baseline (after CLEF07)	3000	Okapi	yes	0.2611
Sinai Filtering (CLEF07)	1000	TF.IDF	yes	0.1343
Sinai Filtering (CLEF07)	1000	Okapi	yes	0.2216
Sinai Filtering (after CLEF07)	1000	Okapi	yes	0.2605
Sinai Filtering (after CLEF07)	3000	Okapi	yes	0.2611

5 Conclusions and Future Work

In this paper we present the new experiments carried out after our second participation in the GeoCLEF 2007 campaign. In GeoCLEF 2007[1] we have introduced a very restrictive system: we have tried to eliminate those documents recovered by the IR subsystem that do not satisfy certain validation rules. However, in GeoCLEF 2006[2] we tried the expansion of queries with entities and thesauri information in order to improve retrieval effectiveness. The main conclusion is that filtering of the documents retrieved works better than the expansion of queries.

The evaluation of the results obtained in these post experiments carried out after the GeoCLEF 2007 showed us the following conclusions:

- The documents filtered by the GR validator subsystem are valid but their positions in the final ranking could be better.
- The Okapi weighting function works better than TF.IDF in the information retrieval subsystem.
- The inclusion of Pseudo-Relevant Feedback (PRF) as parameter for information retrieval process gives better results than the exclusion of it.
- The increasing of the number of documents retrieved by the information retrieval subsystem also improves the results obtained.

For the future, we will try to add more heuristics to the GR validator subsystem making use of the Geonames gazetteer. We will also define more precise rules in the filtering process so that the system obtains better results. Our future work will include the test of a re-indexing method with several relevant

documents validated by the GR validator subsystem and the queries will be run against the new index.

Acknowledgments

This work has been partially supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id.514 granted by University of Jaén.

References

1. Perea-Ortega, J.M., García-Cumbreras, M.A., García-Vega, M., Montejo-Ráez, A.: GEOUJA System. University of Jaén at GEOCLEF 2007. In: Working Notes of the Cross Language Evaluation Forum (CLEF 2007), p. 52 (2007)
2. García-Vega, M., García-Cumbreras, M.A., Ureña-López, L., Perea-Ortega, J.M.: GEOUJA System. The first participation of the University of Jaén at GEOCLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
3. Porter, M.F.: An algorithm for suffix stripping. Program 14, 130–137 (1980)
4. García-Cumbreras, M.A., Ureña-López, L.A., Santiago, F.M., Perea-Ortega, J.M.: BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. LNCS. Springer, Heidelberg (2007)
5. Robertson, S., Walker, S.: Okapi-Keenbow at TREC-8. In: Proceedings of the 8th Text Retrieval Conference TREC-8, pp. 151–162. NIST Special Publication 500-246 (1999)
6. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic query expansion using smart: Trec 3. In: Proceedings of TREC3, pp. 69–80. NIST, Gaithersburg (1995)

TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier

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

Abstract. This paper describes and analyzes the results of our experiments in Geographical Information Retrieval (GIR) in the context of our participation in the CLEF 2007 GeoCLEF Monolingual English task.

Our system uses Linguistic and Geographical Analysis to process topics and document collections. Geographical Document Retrieval is performed with Terrier and Geographical Knowledge Bases.

Our experiments show that Geographical Knowledge Bases can be used to improve the retrieval results of the Terrier state-of-the-art IR system by filtering out non geographically relevant documents.

1 System Description

Our GIR system is a modified version of the system presented in GeoCLEF 2006 [1] with some changes in the Retrieval modes and the Geographical Knowledge Base. The system has four phases performed sequentially: i) a Linguistic and Geographical Analysis of the topics, ii) a thematic Document Retrieval with Terrier (a state-of-the-art search engine that implements relevance feedback and several retrieval models such as: TFIDF, BM25, and Divergence From Randomness), iii) a Geographical Retrieval task with Geographical Knowledge Bases (GKBs), and iv) a Document Filtering phase. In addition, we have developed a toolbox based on Shape Files ¹ for countries, following [2]. A Shape File is a popular geospatial vector data format for geographic information systems software. Shape Files spatially describe geometries: points, polylines, and polygons.

In this paper we focus on the analysis of the experiments at GeoCLEF 2007. For a more detailed description of the system architecture, collection processing, and tuning, consult [3].

2 Experiments

For the GeoCLEF 2007 evaluation we performed a set of five experiments applying geographical knowledge filtering, Relevance Feedback, and different topic

¹ <http://www.esri.com>

tags to an automatic state-of-the-art IR system, resulting in the five runs we submitted to this evaluation. Consult [3] for details.

3 Results

The results of the TALPGeoIR system at the GeoCLEF 2007 Monolingual English task are summarized in Table 1. All runs use Relevance Feedback except run *TalpGeoIRTDN3* that uses Geographical border filtering.

Table 1. TALPGeoIR results at GeoCLEF 2007

Run	Tags	IR System	AvgP.	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	Terrier	0.2711	0.2847	91.23%	593/650
TALPGeoIRTD2	TD	Terrier & GeoKB	0.2850	0.3170	90.30%	587/650
TALPGeoIRTDN1	TDN	Terrier	0.2625	0.2526	93.23%	606/650
TALPGeoIRTDN2	TDN	Terrier & GeoKB	0.2754	0.2895	90.46%	588/650
TALPGeoIRTDN3	TDN	Terrier & GeoKB	0.2787	0.2890	92.61%	602/650

The global results of our runs are good. Four of our five runs are ranked as the first four runs in the GeoCLEF 2007 evaluation task (consult [4] for more details) both considering Mean Average Precision (ranging from 28.50% to 27.11%, next system was scored 26.42%) and R-Precision (ranging from 31.70% to 28.47%, next system was scored 27.23%).

In order to analyse the source of errors we have examined our less reliable topics, i.e. 1) all having a score clearly under the Mean Average Precision for any of our runs (topics 4, 11, 16 and 17) and 2) all having a score close to the Mean Average Precision for more than one run (topics 2, 9, 10, 12, 13, 14, 20, 21, 23 and 24). We reproduce the title of these topics in figure Table 2.

We have used information as the number of relevant documents recovered and the different scores provided by the organisation and found the following main sources of error (we illustrate each case with some examples corresponding to these topics).

1. Failing on properly recognizing toponyms.
 - (a) Sometimes the location term has not been located in our gazetteers due to lack of coverage or different spelling, e.g. “Nagorno-Karabakh” in topic 10.
 - (b) Sometimes there is a problem of segmentation. For instance, “Mediterranean Sea” (13) has been considered a multiword term by our NER and has not been located as so in our gazetteers.
 - (c) Errors from the NERC classifying incorrectly toponyms as persons. For instance “Vila Franca de Xira” has been recognized as a person by our NERC system.
 - (d) Our gazetteers have not recognized “St Paul’s Cathedral”. There is a lack of important facilities in our gazetteers.

Table 2. Less reliable topics GeoCLEF 2007

Num	Topic Title
2	Crime near St Andrews.
4	Damage from acid rain in northern Europe.
9	Meetings of the Andean Community of Nations (CAN).
10	Casualties in fights in Nagorno-Karabakh.
11	Airplane crashes close to Russian cities.
12	OSCE meetings in Eastern Europe.
13	Water quality along coastlines of the Mediterranean Sea.
14	Sport events in the french speaking part of Switzerland.
16	Economy at the Bosphorus.
17	F1 circuits where Ayrton Senna competed in 1994.
20	Tourist attractions in Northern Italy.
21	Social problems in greater Lisbon.
23	Events at St. Paul's Cathedral.
24	Ship traffic around the Portuguese islands.

- (e) Our NERC uses to perform correctly but failed to classify “CAN” (9) as an organization and classified it as a locative and “CAN” was found as a synonym of “Canada” in our gazetteers.
- 2. Failing on properly disambiguating toponyms. In some cases the toponyms have been correctly recovered from the gazetteers but the disambiguation process was wrong. This was the case of “Columbia” (narrative of 9), a typo in the text, that has been located in USA.
- 3. Acronyms have not been expanded for refining queries. For instance, “OSCE” (12) has not been expanded.
- 4. The system did not refined the query with hyponyms. This limited in some cases the coverage. Neither “Crime” (2) nor “Economy” (16) have been refined beyond the examples included in the narrative.
- 5. GEO relations (as in 20) have been properly extracted but are used only in TDN3 run to apply a border filtering algorithm that has been used in 6 topics.
- 6. Sometimes as in (“F1 circuits”, 17) no locative has been found.
- 7. We have failed to attach complementary locative descriptors to the geographic term as in “Russian cities” (11) or “coastlines” (13).

The border filtering algorithm has been used in the following topics of the run TDN3 (topics 2, 8, 16, 19, 21, and 25) applying a configuration of the Terrier IR without query expansion. Compared with the run TDN2 the MAP improves slightly in topics 8, and 25 but drops in topics 2, 16, and 19. The use of border filtering without query expansion seems that does not provide a general improvement neither in MAP nor in recall. On the other side, analyzing the topics that do not use border filtering without query expansion (19 topics) in run TDN3 and comparing them with the same topics in run TDN2, seems that at least in three topics avoiding query expansion has supposed a great improvement

in recall and MAP (topics 1, 3, and 7), only there is a slightly drop in MAP in topics (10 and 20) and a noticeable drop in recall in topic 23.

4 Conclusions

Our approach with both Terrier and a Geographical Knowledge Base shows that applied GKBs can improve some retrieval results of a state-of-the-art IR system. The approach with Terrier and the GeoKB was slightly better in terms of MAP than the one with Terrier alone. The topics with Border Filtering approach applied without Relevance Feedback do not perform a general improvement of the results in MAP and Recall.

As future work we propose the following improvements to the system: i) the resolution of geographical ambiguity problems applying toponym resolution algorithms, ii) use Terrier with the Divergence From Randomness algorithm instead of TFIDF, iii) the improvement and evaluation of the Shape Files toolbox and the Border Filtering algorithm.

Acknowledgments

This work has been supported by the Spanish Research Dept. (TEXT-MESS, TIN2006-15265-C06-05). 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. Ferrés, D., Rodríguez, H.: TALP at GeoCLEF 2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 962–969. Springer, Heidelberg (2007)
2. Pouliquen, B., Steinberger, R., Ignat, C., Groeve, T.D.: Geographical Information Recognition and Visualization in Texts Written in Various Languages. In: SAC 2004, pp. 1051–1058. ACM Press, New York (2004)
3. Ferrés, D., Rodríguez, H.: TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (September 2007)
4. Mandl, T., Gey, F., Nunzio, G.D., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xie, X.: GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (September 2007)

TALP at GeoQuery 2007: Linguistic and Geographical Analysis for Query Parsing

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

Abstract. This paper describes our experiments and analysis of the results of our participation in the Geographical Query Parsing pilot-task for English at GeoCLEF 2007. The system uses deep linguistic analysis and Geographical Knowledge to perform the task.

1 Introduction

The Query Parsing task (GeoQuery) is a pilot task proposed in GeoCLEF 2007. It consists on five subtasks: i) detect whether the query is geographic or no, ii) extract the WHERE component of the query, iii) extract the GEO-RELATION (from a set of predefined types) if present, iv) extract the WHAT component of the query and classify it as Map, Yellow Page or Information types, v) extract the coordinates (Lat-Long) of the WHERE component.

Our system uses some modules of a Geographical Information Retrieval system presented at GeoCLEF 2006 [1] and modified for GeoCLEF 2007 [2].

In this paper we present the overall architecture of our Geographical Query Parsing system, our experiments, analysis of the results and conclusions in the context of the GeoCLEF's 2007 GeoQuery pilot task.

The paper focuses on the analysis of the results. See more details about the system implementation in [2].

2 System Description

For each of the 800,000 queries our systems performs a two steps process:

1. Linguistic and Geographical Analysis of the topic.
2. Query Classification and Information Extraction.

2.1 Linguistic and Geographical Analysis of the Topic

The queries are first linguistically processed by our Natural Language processors resulting in the following structures: i) Sent, which provides lexical information for each word: form, lemma, POS tag, semantic class of NE, list of WN synsets.

ii) Sint, composed of two lists, one recording the syntactic constituent structure of the query and the other collecting the information of dependencies and other relations between these components. iii) Environment, the environment represents the semantic relations that hold between the different components identified in the topic text (as shown in Figure 1). See [3] for details on the way of building these structures

The Geographical Analysis is applied to the Named Entities from the queries that have been classified as Location or Organization by the NERC module. A Geographical Thesaurus is used to extract geographical information about these Name Entities. This component has been built joining four gazetteers that contain entries with places and their geographical class, coordinates, and other information: i) GEOnet Names Server (GNS), ii) Geographic Names Information System (GNIS), iii) GeoWorldMap World Gazetteer , iv) World Gazetteer. A subset of the most important features from this thesaurus has been manually set using 46,132 places (including all kind of geographical features: countries, cities, rivers, states,...). This subset of important features has been used to decide if the query is geographical or not geographical. See an example of the Linguistic and Geographical Analysis in Figure 1.

2.2 Query Classification and Information Extraction

The Query Classification task is performed through the following steps:

- Over the sint structure, a DCG like grammar consisting of about 30 rules developed manually from the sample of GeoQuery and the set of queries of GeoCLEF 2006, is applied for obtaining the topics and geographic mentions (including relations if present) occurring in the query. A set of features (consultive operations over chunks or tokens and predicates on the corresponding sent structures) is used by the grammar (see [2] for details).
- Finally from the result of step 2 several rule-sets are in charge of extracting: i) LOCAL, ii) WHAT and WHAT-TYPE, iii) WHERE and GEO-RELATION, and iv) LAT-LONG data. So, there are four rule sets with a total of 25 rules.

3 Experiments and Results

We performed only one experiment for the GeoQuery2007 data set. The global results of our run for the local query were 0.222 Precision, 0.249 Recall, and 0.235 of F1. Our system was ranked the third from 6 participants.

Query: "Discount Airline Tickets to Brazil"
 Semantic: [entity(3),mod(3,1),quality(1),mod(3,2),entity(2),i-en_country(5)]
 Linguistic: Brazil Brazil NNP Location
 Geographical: America@@South_America@@Brazil@@-10.0_-55.0
 Feature type: administrative_areas@@political_areas@@countries

Fig. 1. Semantic and Geographical Content of GQ-38

In order to analyse the source of errors we have implemented the evaluation criteria described in [4]. The confusion matrices for LOCAL and WHAT-TYPE for the 500 queries evaluated are presented in tables 1 and 2. The number of errors have been 99 for LOCAL, 126 for WHAT, 179 for WHAT-TYPE, 41 for GEO-RELATION, 122 for WHERE, giving a total of 315 queries with one or more errors and 185 correctly answered.

Table 1. Confusion matrix for LOCAL

	NO	YES
NO	122	31
YES	68	279

Table 2. Confusion matrix for WHAT-TYPE

	Map	Information	Yellow Pages
Map	4	4	10
Information	44	39	121
Yellow Pages	38	28	212

We will focus on the most problematic figures:

1. Queries not recognized as LOCAL (31) by our system. Clearly this case corresponds to the different coverage of our gazetteers and those used by the evaluators. Some frequent errors can be classified as follows:
 - Some errors simply correspond to lack of coverage of our gazetteers (as “cape may”).
 - Some of these errors could be recovered using the context (as “Gila County”).
 - Sometimes the query has been considered as LOCAL because it corresponds to an address (street, place and so). We have not considered these kinds of locations (as “caribbean joe”).
 - There some typos (as “nuernberg”, placed in Germany and probably corresponding to “Nuremberg”).
 - Some cases correspond to misspellings of Spanish words (as “Cercanías” considered erroneously as a toponym is Spain).
2. Queries we have improperly considered as LOCAL (68): i) in some cases (30%) it seems that our gazetteers have a higher coverage. ii) Other queries correspond to Named Entities probably not present in our gazetteers but erroneously classified as location by our NERC (as “Hitachi” or “Sala”).
3. From table 2 the following problematic cases arise: i) confusion between “Map” and “Information” or “Yellow Pages”. Most of the errors correspond to a lack of a rule that assigns “Map” to the queries consisting on only a locative (as “Coronado, San Diego”). Sometimes it is due to an only partial recognizing of the locative. ii) the confusion between “Information” and

“Yellow Pages” is problematic as [4] point out. There is no clear trends on the typification of the errors. Besides, we have used “Yellow Pages” as our default class when no classification rule can be applied. Obviously a more precise classification is needed.

4 Conclusions

Our system for the GeoCLEF’s 2007 GeoQuery pilot task is based on a deep linguistic and geographical knowledge analysis. Our analysis of the results show that the selection of a subset of the most important features to create a gazetteer of only the most important places implies a lost of coverage and thus missing geographical places and classifying queries as non-local. As a future work we plan to apply more sophisticated ways to create subsets of geographically relevant places that could appear in web search queries.

Acknowledgments

This work has been supported by the Spanish Research Dept. (TEXT-MESS, TIN2006-15265-C06-05). 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. 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, 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. 947–955. Springer, Heidelberg (2006)
2. Ferrés, D., Rodríguez, H.: TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2007)
3. Ferrés, D., Kanaan, S., Ageno, A., González, E., Rodríguez, H., Surdeanu, M., Turmo, J.: The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 557–568. Springer, Heidelberg (2005)
4. Li, Z., Wang, C., Xie, X., Ma, W.Y.: Query parsing task for GeoCLEF2007 report. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop, Budapest, Hungary (September 2007)

Applying Geo-feedback to Geographic Information Retrieval

Mirna Adriani and Nasikhin

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
mirna@cs.ui.ac.id, nasikhin@mhs.ui.edu

Abstract. In this paper we identify location names that appear in queries written in Indonesian using geographic gazetteer. We built the gazetteer by collecting geographic information from a number of geographic resources. We translated an Indonesian query set into English using a machine translation technique. We also made an attempt to improve the retrieval effectiveness using a query expansion technique. The result shows that identifying locations in the queries and applying the query expansion technique can help improve the retrieval effectiveness for certain queries.

1 Introduction

As our participation in the Geographical Information Retrieval of the Cross Language Evaluation Forum (CLEF 2007) task, i.e., for Indonesian-English, we needed to use language resources to translate Indonesian queries into English. We learned from our previous work [1] that freely available dictionaries on the Internet could not correctly translate many Indonesian terms, as their vocabulary was very limited. Luckily we found a machine translation tool available on the Internet that could help translate the Indonesian queries into English. However, GIR focuses on identifying geographical names [2] that appear in the queries, so we also needed to work on translating the location names from English to Indonesian.

2 The Process of Identifying Location Names

There are many resources that contain geographical information available on the Internet. We made use of the *gazetteer* to build a location hierarchy map. The location hierarchy was built by extracting the names of countries, provinces, etc. from the *gazetteer*. For each location, information about other locations within the area that it covers was added, such as cities under a province, etc. We obtained the needed geographical information from Geonames¹ and Wikipedia².

¹ <http://www.geonames.org/>

² <http://id.wikipedia.org/>

We extracted the names of provinces, their capital cities, the names of mountains, seas etc. from Geonames and its translation in Bahasa Indonesia from Wikipedia. Each location has all their alternate names in both English and Bahasa Indonesia. If one location name appears in a query or document, it will be looked up in the gazetteer to find its associated locations that can be used as terms for searching for or indexing the document.

Most documents in the collection contain information about the location of events. For each document, we identified the location where the event mentioned in the document occurred, and added the location information into the document's index entry. For documents that contain more than one location, we choose the location that has the highest frequency in the document. If there is more than one location with the same highest frequency then a location is selected randomly among such locations.

To process the geographical locations further, we identify words that are related to a location name such as *in the (north/south/...) of*, *in the border of*, *around* etc. Then we include all location names that fall inside a boxline surrounding the location (city, country etc.) The boxline borders are at certain distance north, south, east, and west of a location.

2.1 Query Expansion Technique

Adding translated queries with relevant terms (query expansion) has been shown to improve CLIR effectiveness [3], [4]. One of the query expansion techniques is called the *pseudo relevance feedback* [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 applied this technique in this work. To choose the relevant terms from the top ranked documents, we used the $tf*idf$ term weighting formula [5]. We added a number of terms with 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 that came with the query topics. The query translation process was performed fully automatic using a machine translation technique. The machine translation technique translates the Indonesian queries into English using Toggletext³, a machine translation tool that is available on the Internet. Any location names that appear on the query will be identified and used in searching documents. We then applied a pseudo relevance-feedback query-expansion technique to the queries that were translated using the machine translation above. Besides adding terms, we also add location names only that appear on the top documents. In these experiments, we used the Lemur⁴ information retrieval system which is based on the language model to index and retrieve the documents.

³ <http://www.toggletext.com/>

⁴ See <http://www.lemurproject.org/>

4 Results

Our work focused on the bilingual task using Indonesian queries to retrieve documents in the English collections. Table 1 shows the result of our experiments.

Table 1. Average retrieval precision of the monolingual runs of the title, their translation queries, and the use of the geographic identification and query expansion on the translated queries

Task	Monolingual	% Change
Title	0.1767	-
Title (translation)	0.1417	-19.80%
Title (Geoprocessing)	0.1736	-1.75%
Title (Geoprocessing + Geofeedback: 10 docs, 5 locs)	0.1389	-21.39%
Title (Geoprocessing + Pseudofeedback: 5 docs, 5 terms)	0.1936	+9.56%

Table 2. Average retrieval precision of the monolingual runs of the combination of title and description topics, their translation queries, and the use of the geographic identification and query expansion on the translated queries

Task	Monolingual	% Change
Title + Description	0.1979	-
Title + Description (translation)	0.1812	-8.43%
Title + Description (Geoprocessing)	0.2096	+5.91%
Title + Description (Geoprocess + Geofeedback: 5 docs, 5 locs)	0.2091	+5.65%
Title + Description (Geoprocess + Pseudofeedback: 10 docs, 5 terms)	0.1939	-2.02%

The retrieval performance of the title-based translation queries dropped 19.80% below that of the equivalent monolingual retrieval (see Table 1). The retrieval performance of location identification process on the queries dropped 1.75% below that of the equivalent monolingual queries. Expanding the queries by adding geographic location from the top documents to the translated queries decreases the retrieval performance by 21.39%. However, adding terms that appear on the top documents on the translated queries improve the retrieval performance by 9.56%.

The retrieval performance of the combination of title and description queries that is translated by machine translation dropped 8.43% below that of the equivalent monolingual retrieval (see Table 2). The identification of the location on the queries improves the average precision 5.91%. Expanding the queries by adding geographic locations that appear from the top documents increases the average precision by

5.65%. However, adding terms from the top documents decreases the average precision by 2.02%.

5 Summary

Our results demonstrate that identifying location on the queries can have a positive and negative effect on the queries. The query expansion technique that was applied to the queries by adding more terms and location names also produced mixed results. For the title queries, the query expansion had a positive impact when the combination of terms and location names were added to the queries. However, the same situation did not work for the combination of title and description queries. It had a positive impact only when the queries were added with terms or location names only. We still need to study further on the effect of location identification because the decrease in retrieval performance was not only caused by the failure in identifying the correct location names but also the failure in translating the words and location names in the queries from one language to another.

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. Overell, S.E., Ruger, S.: Identifying and Grounding Descriptions of Places. In: Proceedings of the Geographic Information Retrieval workshop (GIR 2006), Seattle, USA (2006)
3. Buscaldi, D., Rosso, P., Garcia, P.P.: WordNet-based Index Terms Expansion for Geographical Information Retrieval. In: Nardi, A., Peters, C., Vicedo, J.L. (eds.) Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop (CLEF 2006) (2006)
4. Larson, R.R.: Cheshire II at GeoCLEF: Fusion and Query Expansion for GIR. In: Peters, C. (ed.) Cross Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF 2005) (2005)
5. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)

Exploring LDA-Based Document Model for Geographic Information Retrieval

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. Latent Dirichlet Allocation (LDA) model, a formal generative model, has been used to improve ad-hoc information retrieval recently. However, its feasibility and effectiveness for geographic information retrieval has not been explored. This paper proposes an LDA-based document model to improve geographic information retrieval by inheriting the LDA model with text retrieval model. The proposed model has been evaluated on GeoCLEF2007 collection. This is a part of the experiments of Columbus Project of Microsoft Research Asia (MSRA) in GeoCLEF2007 (a cross-language geographical retrieval track which is part of Cross Language Evaluation Forum). This is the second time we participate in this event. Since the queries in GeoCLEF2007 are similar to those in GeoCLEF2006, we leverage most of the methods that we used in GeoCLEF2006, including MSRAWhitelist, MSRAExpansion, MSRALocation and MSRAText approaches. The difference is that MSRManual approach is not included this time, and we use MSRALDA instead. The results show that the application of LDA model in GeoCLEF monolingual English task performs stably but needs to be further explored.

Keywords: Geographic information retrieval, System design, Latent Dirichlet Allocation, Evaluation.

1 Introduction

In general web search and mining, location information is usually discarded. However, people need to deal with locations all the time, such as dining, traveling and shopping. GeoCLEF [1] aims at providing necessary platform for evaluating geographic information retrieval (GIR) systems. We have participated in GeoCLEF2006, and in [2] we proposed several query processing methods, including manual expansion, pseudo-feedback, whitelist-method and location extraction. This is the second time we participate in GeoCLEF event.

Incorporating topic models with search algorithms has a long history in information retrieval. For example, Latent Semantic Indexing (LSI) technique [3] was introduced in 1990. Probabilistic Latent Semantic Indexing model (pLSI) model [4] represents

documents as a mixture of topics by using a latent variable and outperforms LSI in small corpus, e.g. thousands of documents. To overcome the overfitting problem, Latent Dirichlet Allocation (LDA) model [5] is proposed by using a probabilistic graphical model. Compared with pLSI, LDA processes fully consistent generative semantics by treating the topic mixture distribution as a K -parameter hidden variable rather than a large set of individual parameters which are explicitly linked to the training set. LDA has shown its promising effectiveness in ad-hoc retrieval in [6]. However, its feasibility and effectiveness remain unknown for geographic information retrieval. In this paper, we proposed an LDA-based document retrieval model to improve the search quality for GIR and evaluated it in GeoCLEF2007.

2 Geographic Information Retrieval System: An Overview

Fig. 1 shows the system flow of our GIR system used in GeoCLEF2007. Our geographic information retrieval system is mainly composed of Geo-knowledge base (GKB), location extraction module, Geo-focus detection module, query-processing module, Geo-indexing module and Geo-ranking module. More details about these modules can be found in [7]. In general, we index the collections provided by the organizers using the index schemes in [8] in offline-phase. In the query processing module, the topics are translated into our input format based on several approaches including automatic extraction, pseudo-feedback and manual expansion. Finally, we rank the results based on different ranking models which we will discuss here in detail.

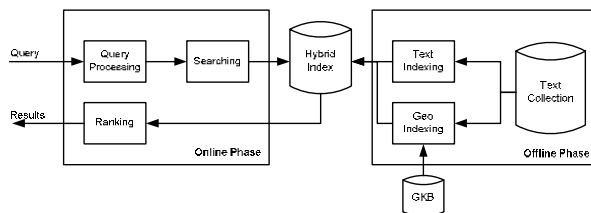


Fig. 1. Architecture of our GIR system

3 Geo-ranking Module

We modified the IREngine, developed by MSRA, by integrating the Geo-ranking module as our basic search engine. We totally adopted three kinds of ranking models: 1) pure textual ranking model. Its basic ranking model is vector model; 2) Geo-based Model; 3) LDA-based Model. They will be described in the following sections.

3.1 Geo-based Model

In Geo-based model, we retrieve a document list according to Geo-relevance from the Geo-index first. That is, for the focus-index which utilizes the inverted index to store all the explicit and implicit locations of documents, the matched document identifier (docID) list can be retrieved by looking up the query locations in the inverted index.

For the grid-index which divides the surface of the Earth into 1000×2000 grids, we can get the docID list by looking up the grids in which the query locations fall into. We also retrieve a list of documents relevant to the textual terms, and merge the two lists to get the final results. Finally, we use a combined ranking function $R_{combined} = R_{text} \times \lambda + R_{geo} \times (1 - \lambda)$, where R_{text} is the textual relevance score and R_{geo} is the Geo-relevance score, to re-rank the results. Experiments show that textual relevance scores should be weighted higher than Geo-relevance scores ($\lambda = 0.8$ in our experiments).

3.2 LDA-Based Model

In LDA-based model, we explored the Latent Dirichlet Allocation model in our Geo-CLEF2007 experiments. Latent Dirichlet Allocation (LDA) model is a semantically consistent topic model. In LDA, the topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all documents. The graphical model of LDA is shown in Fig. 2.

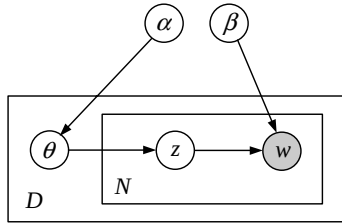


Fig. 2. Graphical representation of LDA model

The generative process for LDA can be stated as follows:

For each text document d :

1. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
2. For each word w_n in document d :
 - a. Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
 - b. Choose a word $w_n \sim p(w_n | z_n, \beta)$, which is a topic-specific multinomial probability distribution.

Thus, the likelihood of generating a corpus $D = \{d_1, d_2, \dots, d_M\}$ is:

$$P(D | \alpha, \beta) = \prod_{m=1}^M \int p(\theta_m | \alpha) \left(\prod_{n=1}^{N_m} \sum_{z_{mn}} p(z_{mn} | \theta_m) p(w_{mn} | z_{mn}, \beta) \right) d\theta_m$$

In [6], Xing et al. discussed the application of LDA in ad hoc retrieval. We use the similar approach for our geographic information retrieval task in GeoCLEF2007, which allows us to compute a probability of a query given a document using LDA model. That is each document is scored by the likelihood of its model generating a query Q ,

$$P(Q|d) = \prod_{q \in Q} P(q|d)$$

where d is a document model, Q is the query and q is a query term in Q . $P(Q|d)$ is the likelihood of the document model generating the query terms under the “bag-of-words” assumption that terms are independent given the documents. In our experiment, we use LDA model as the document model.

After we computed the $P_{lda}(Q|d)$, we selected the top 1000 documents with the highest $P_{lda}(Q|d)$ for each query. We also use our text search engine to retrieve top 1000 documents respectively. Then we merged these two document-lists. If one document in both of the list, we used a combined score function $R_{combined} = R_{text} \times \lambda + P_{lda} \times (1 - \lambda)$, where R_{text} is the textual relevance score and P_{lda} is the LDA model probability (here we set $\lambda = 0.5$). Both scores are normalized. Otherwise, we computed a new score for the document by multiplying a decay factor 0.5. Finally we re-ranked all these documents based on the new scores and selected the top 1000 ones as result.

4 Monolingual GeoCLEF Experiments (English - English)

In Table 1, we show all the five runs submitted to GeoCLEF. The topics in GeoCLEF mainly consist of three elements: “title”, “desc” and “narr”. “Title” gives the search goal of topic, “desc” shows the standard for the related results and “narr” contains more description of the information requested defined by the topic, including specifics about the geography of the topic such as a list of desired cities, states, countries or latitudes and longitudess. When the topic field is “Title”, we just use the title element of the topic to generate the query of the run. When the topic field is “Title + Description”, this means that the title and desc elements are both used in the run. When the topic field is “Title + Description + Narrative”, this means that title, desc and narr elements 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
MSRALDA	Title	1
MSRAWhiteList	Title + Description	2
MSRAExpansion	Title + Description	3
MSRALocation	Title	4
MSRAText	Title + Description + Narrative	5

In MSRALDA, we used the title elements to generate the queries. Then we used the LDA-based model to select 1000 documents for each query. In MSRAWhiteList, we used the Title and Desc elements of the topics to generate the queries. For some special queries, e.g. “Scottish Islands”, “coastlines of the Mediterranean Sea”, we cannot get the exact locations directly from our gazetteer, so we utilized the GKB to get the corresponding Geo-entities. Then we can make a whitelist manually for the

Geo-terms of these queries. In MSRAExpansion, we generated the queries with title and desc elements of the topics. Different from MSRAWhiteList, the queries were automatically expanded based on the pseudo-feedback technique. First we used the original queries to search the corpus. Then we extracted the locations from the returned documents and calculated the times each location appears in the documents. Finally we got the top 10 most frequent location names and combined them with the original Geo-terms in the queries. In MSRALocation, we used the 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 used our location extraction module to extract the locations automatically from the queries. Geo-based model is used in MSRAWhiteList, MSRAExpansion and MSRALocation to rank the results. In MSRAText, we generated the queries with title, desc and narr elements of the topics. We utilized our pure text search engine “IREngine” to process the queries.

5 Results and Discussions

Fig. 3 and Table 2 show the results of MSRA Columbus on the GeoCLEF monolingual English task. MSRAText run achieves the best precision in our results. The precision of MSRALDA run decreases after combing the LDA model with the pure text model. As same as GeoCLEF2006, the performance of MSRAExpansion is the lowest among the five runs, because many unrelated locations are added to new topics after pseudo-feedback for some topics.

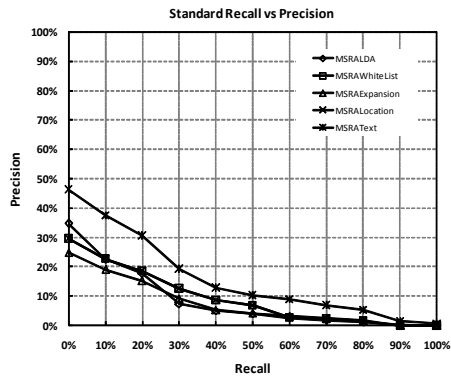


Fig. 3. Standard recall levels vs mean interpolated precision for all five runs

From Table 2, we can see that MSRALDA drops the performance significantly compared with MSRAText by about 7.6% in mean average precision (MAP). This indicates that linearly combining LDA model with text model does not work well. The reason may be that linear combination of text score and LDA probability is not a good choice. Text score is computed based on some kind of heuristic function and LDA probability is generated by a topic model that learned from the training set. Directly combining two values from different spaces may bring down the results

Table 2. MAP & Standard Deviation for five runs

RUN-ID	MAP	Standard Deviation
MSRALDA	7.51%	0.090
MSRAWhiteList	8.61%	0.145
MSRAExpansion	7.01%	0.134
MSRALocation	8.61%	0.145
MSRAText	15.19%	0.197

Table 3. Topic by Topic Comparision in Three Metrics

Topic	Mean of All Runs			MSRAText			MSRALDA		
	#Rel-ret	Ave-Pre	R-Pre	#Rel-ret	Ave-Pre	R-Pre	#Rel-ret	Ave-Pre	R-Pre
51	91.2	0.4794	0.4773	98.0	0.5422	0.5536	99.0	0.3375	0.2946
52	3.0	0.0135	0.0292	0	0	0	0	0	0
53	41.3	0.114	0.1595	39.0	0.0984	0.1563	43.0	0.1597	0.2656
54	4.6	0.0597	0.0863	4.0	0.0085	0	3.0	0.0272	0
55	11.9	0.1263	0.1462	6.0	0.0060	0	10.0	0.1004	0.1875
56	5.1	0.1416	0.1651	3.0	0.1466	0.2500	3.0	0.0071	0
57	11.7	0.1736	0.2201	14.0	0.3403	0.4000	14.0	0.0792	0.0667
58	4.3	0.0569	0.0597	5.0	0.1481	0.1667	5.0	0.0531	0.1667
59	0	0	0	0	0	0	0	0	0
60	22.0	0.524	0.5094	28.0	0.7791	0.7333	11.0	0.2018	0.2000
61	15.1	0.1249	0.1458	10.0	0.0468	0.0909	11.0	0.1700	0.2273
62	2.7	0.1375	0.1132	4.0	0.0491	0	1.0	0.0003	0
63	0	0	0	0	0	0	0	0	0
64	7.1	0.028	0.0411	2.0	0.0003	0	3.0	0.0031	0
65	66.8	0.2134	0.2794	59.0	0.1407	0.2581	74.0	0.1268	0.1398
66	1.79	0.1663	0.1761	1.0	0.0021	0	0	0	0
67	12.9	0.2863	0.2719	17.0	0.3663	0.2941	17.0	0.2390	0.2353
68	104.3	0.4713	0.4963	65.0	0.2566	0.3359	69.0	0.1243	0.2595
69	12.7	0.1166	0.1608	11.0	0.0582	0.1429	12.0	0.1039	0.1905
70	15.9	0.0587	0.0839	20.0	0.0340	0.0690	22.0	0.0695	0.1379
71	0	0	0	0	0	0	0	0	0
72	11.1	0.3728	0.4043	5.0	0.2737	0.2857	6.0	0.0455	0.0714
73	4.6	0.0462	0.0323	1.0	0.0021	0	1.0	0.0019	0
74	2.1	0.2611	0.2201	2.0	0.3356	0.3333	2.0	0.0078	0
75	13.8	0.2801	0.3104	14.0	0.1631	0.2000	12.0	0.0211	0.0500

quality. Another reason may be that we have not tuned the parameters to be the best. And we can see that the standard deviation of MSRALDA is just 0.09, lower than MSRAText. This indicates that MSRAText performs badly in some cases while MSRALDA performs more stably.

MSRAWhiteList and MSRALocation achieve similar MAP with each other, about 8.6%. Their MAPs are much lower than MSRAText by about 6.5% and just little better than MSRAExpansion. Different from the results of GeoCLEF2006, automatic location extraction and manual expansion don't bring improvements.

Table 3 shows the performance of MSRAText and MSRALDA, and the mean performance of all the runs in Monolingual (EN-EN) task for the 25 topics in three metrics, including number of relevant-retrieved documents (#Rel-ret), average precision (Ave-Pre) and R-precision (R-Pre) which are defined below.

$$\text{Average precision} = \frac{\sum_{r=1}^N (P(r) \times \text{rel}(r))}{\text{number of relevant documents}}$$

where r is the rank, N is the number of retrieved documents, $\text{rel}()$ is a binary function on the relevance of a given rank, and $P()$ is precision at a given cut-off rank. R -precision is the precision at R where R is the number of relevant documents in the collection for the query.

Though the MAP of MSRALDA is lower than MSRAText, MSRALDA still outperforms MSRAText in the topic 53, 54, 55, 61, 64, 69 and 70 in the average precision metric. For example, for the topic 53 “Scientific research at east coast Scottish Universities”, MSRAText just retrieves 39 relevant documents, while MSRALDA retrieves 43 relevant ones (The number of relevant documents is 64) and its average precision increases from 0.0984 to 0.1597. For the topic 70 “Tourist attractions in Northern Italy”, MSRAText retrieves 20 relevant documents and MSRALDA retrieves 22 (The number of relevant documents is 29) and its precision increases from 0.0340 to 0.0695. We conclude that more related documents are found after combining LDA model because LDA model can find the relationship between words and documents even though the words do not exist in the documents. Interestingly, though MSRALDA retrieves more relevant documents, it brings down the precision for some topics. For instance, for the topic 65, MSRALDA retrieves 74 relevant documents, 15 ones more than MSRAText. But its average precision drops about 1.5% and R -precision drops about 12%. The reason is that the ranking of relevant documents is not correct though MSRALDA retrieves more relevant documents.

The average precision of the topic 51, 56, 57, 58, 60, 62, 66, 67, 68, 72, 73, 74 and 75 in MSRALDA are lower than MSRAText. In topic 60, 62, 66, 75, MSRALDA retrieves less relevant documents than MSRAText, so its precision drops. However, in other topics the relevant documents in MSRALDA are not less than MSRAText. The reason for the low precision in this situation is also due to the incorrect ranking of documents.

Compare with the mean performance of all runs, MSRALDA only improves in topic 53, 61 and 70 in the average precision metric.

6 Conclusion

We conclude that the application of LDA model in GeoCLEF monolingual English task performs stably but needs to be further explored, especially the ranking function. Another conclusion is that automatic location extraction from the topics does not improve the retrieval performance, even decrease it sometimes. The third conclusion is the same as last year. That is automatic query expansion by pseudo-feedback weakens the performance because the topics are too hard to be handled and many unrelated locations are added to new topics. Obviously, we still need to improve the system in many aspects, such as query processing, Geo-indexing and Geo-ranking.

References

1. GeoCLEF2007, <http://ir.shef.ac.uk/geoclef/>
2. Zhi-Sheng, L., Chong, W., Xing, X., Xufa, W., Wei-Ying, M.: MSRA Columbus at GeoCLEF 2006. In: Peters, C., et al. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 926–929. Springer, Heidelberg (2007)

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(6), 391–407 (1990)
4. Hofmann, T.: Probabilistic latent semantic indexing. In: *The 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pp. 50–57. ACM Press, New York (1999)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *Journal of Machine Learning Research*, 993–1022 (2003)
6. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: *The 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pp. 178–185. ACM Press, New York (2006)
7. Zhi-Sheng, L., Chong, W., Xing, X., Xufa, W., Wei-Ying, M.: MSRA Columbus at Geo-CLEF 2007. In: *Cross-Language Evaluation Forum: Geographic Information Retrieval Track*, working notes, Budapest, Hungary (2007)
8. Zhi-Sheng, L., Chong, W., Xing, X., Xufa, W., Wei-Ying, M.: Indexing implicit locations for geographical information retrieval. In: *The 3rd Workshop on Geographical Information Retrieval*, pp. 68–70 (2006)

Mono-and Crosslingual Retrieval Experiments with Spatial Restrictions at GeoCLEF 2007

Ralph K lle, Ben Heuwing, Thomas Mandl, and Christa Womser-Hacker

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

Abstract. The participation of the University of Hildesheim focused on the monolingual German and English tasks of GeoCLEF 2007. Based on the results of GeoCLEF 2005 and GeoCLEF 2006, the weighting and expansion of geographic Named Entities (NE) and Blind Relevance Feedback (BRF) were combined and an improved model for German Named Entity Recognition (NER) was evaluated. Post submission experiments are also presented. A topic analysis revealed a wide spread of MAP values with high standard deviation values. Therefore further development will lie in the field of topic-adaptive systems.

1 Introduction

Retrieval of documents which fulfil a spatial requirement is an important task for retrieval systems. Such geographic information retrieval systems are evaluated within the GeoCLEF track at CLEF [1]. Our experiments expanded an ad-hoc system to allow geographic queries. Based on the participation in GeoCLEF 2006 and some post experiments [2], we again adopted a (blind) relevance feedback approach which focuses on named geographic entities. To improve the Named Entity Recognition (NER) for German entities we used an optimized model based on the NEGRA¹ corpus for training. The results (compared to GeoCLEF 2006) did not improve as much as expected, so a topic analysis was performed to identify strategies to adopt the retrieval engine to this kind of topics.

2 Geographic Retrieval System

The system we augmented for this experimentation with (geographic) NEs in GIR is based on a retrieval system applied to ad-hoc retrieval in previous CLEF campaigns [3].

Apache Lucene² is the backbone system for stemming, indexing and searching [2]. Named Entity Recognition was carried out with the open source machine learning

¹ <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html>

² <http://lucene.apache.org/java/>

tool LingPipe³, which identifies named entities and classifies them into the categories Person, Organization, Location and Miscellaneous according to a trained statistical model. During the indexing process, NER leads to an additional index field which contains named geographic entities and can be queried during the search process by weighting and adding these named geographic entities to the blind relevance feedback process.

3 Post Submitted Runs and Analysis

After experimentation with the GeoCLEF data of 2006, we submitted runs differing in parameters and query processing steps.

Run descriptions and results measured as Mean Average Precision (MAP) are shown in Table 1 for submitted monolingual runs and in Table 2 for the corresponding results with the training topics of 2006.

Table 1. Results monolingual runs (nm = new NER-Model)

Run	Language	Narrative	BRF (weight-docs-terms)	Geo-NE's (weight-docs-terms)	MAP
HiMoDeBase	German		0.5-5-25	-	0,2019
HiMoDeNe2	German		0.2-5-25	0.2-30-40 (nm)	0,1953
HiMoDeNe2Na	German	x	0.2-5-25	0.15-30-60 (nm)	0,2067
HiMoDeNe3	German		0.2-5-25	1.0-10-4 (nm)	0,1795
HiMoEnBase	English		0.5-5-25	-	0.1405
HiMoEnNe	English		0.2-5-25	0.5-5-20	0.1535
HiMoEnNaNe	English	x	0.2-5-25	0.5-5-20	0.1497
HiMoEnNe2	English		0.2-5-25	2-10-3	0,1268

With the training topics of 2006, the best results were achieved by expanding the query with 40 geographic terms from the best 30 documents giving each a relative weight of 0.2 compared to the rest of the query (for German) and using 20 terms from the top 5 documents with a relative weight of 0.5 for English (Table 2). With English topics this holds true for the submitted runs, however for German topics the base run without NER performed best (Table 1).

The poor results for the English topics indicate that the topics are more difficult (concerning our retrieval system) for 2007. With the German results remaining on almost the same level, the optimized NER-model for German seems to improve retrieval quality.

Summing up, we could not find a substantial positive impact of additional geographic information, but the effect of investment in optimizing the Geo-NE model seems to be positive. Post submission runs confirmed these impressions, but no significant improvements were found in changing parameters concerning the BRF.

³ <http://www.alias-i.com/lingpipe>

Table 2. Results for training topics (of 2006) of monolingual runs (nm = new NER-Model / om = old NER-Model)

Run	Language	Narrative	BRF (weight-docs-terms)	Geo-NE's (weight-docs-terms)	MAP
HiMoDeBase	German		0.5-5-25	-	0.1722
HiMoDeNe1	German		0.2-5-25	0.2-30-40 (om)	0.1811
HiMoDeNe2	German		0.2-5-25	0.2-30-40 (nm)	0.1963
HiMoDeNe2Na	German	x	0.2-5-25	0.15-30-60 (nm)	0.2013
HiMoDeNe3	German		0.2-5-25	1.0-10-4 (nm)	0.1811
HiMoEnBase	English		0.5-5-25	-	0.1893
HiMoEnNe	English		0.2-5-25	0.5-5-20	0.1966
HiMoEnNaNe	English	x	0.2-5-25	0.5-5-20	0.1946
HiMoEnNe2	English		0.2-5-25	2-10-3	0.1795

4 Topic Analysis

Our topic analysis was divided into two parts: at first we performed a statistical analysis to identify topics which were difficult for all systems and especially for our systems in Hildesheim. Based on the results we carried out an intellectual analysis regarding those topics which performed good resp. bad using different systems.

The first results of the statistical analysis are presented in Table 3. We restricted our analysis on those types of runs which we took part in: monolingual German (DE) and monolingual English (EN). The systems performed slightly better in German (25.76%; 2006: 22.29%) and slightly worse in English (28.5%; 2006: 30.34%) [4], but below any statistical significance.

Table 3. Differences at Standard Deviation of MAP concerning Topics and Systems

Run Type	No. of Participants	No. of Runs	MAP top System	Max. Diff. Stand. Dev. Topics (absolute)	Max. Diff. Stand. Dev. Systems (absolute)
Monoling. DE	4	16	0.2576	0.2320	0.0929
Monoling. EN	11	53	0.2850	0.3118	0.2507

We analyzed the differences between the maximum and the minimum values of the standard deviation regarding the topics on one hand and the systems on the other hand. This shows again that similar to GeoCLEF 2006 the topics have much more influence on the retrieval quality than the systems [2]. There is a special situation for the English tasks, because the worst eight runs resp. systems performed very badly with very small standard deviation values. If these were left out of the calculation, the maximum difference of the standard deviation regarding the systems would be 0.14 instead of 0.25 in Table 3, whereas the maximum difference regarding the topics decreased comparatively slight to 0.24 (from 0.31). In this case the English and the German values are converging.

For a more detailed analysis, the German topics and the four submitted runs were considered. Regarding the absolute values of MAP over all topics, there is a spread of 0.7854 to 0. Taking the best five and the worst five topics away, there remains a difference of 0.2386 (absolute) between the MAPs, which proofs again the influence of the topics on the retrieval results.

As presented in Table 1, the four Hildesheim systems resp. runs performed at an average MAP of 0.1958, the best at 0.2067 (with BRF and Geo-NE, narrative), the worst at 0.1795% (with BRF and Geo-NE, without narrative). Again, only a small difference between the systems can be observed. But of course, there are big differences regarding the relation between the position of every system and the individual topics.

If we performed at an average for the five worst topics (for Hildesheim: 51, 52, 53, 57, 70)⁴, this would improve the performance by almost 5% (absolute, from 0.1958 to 0.2425, relative 24%). Assuming we used all four systems simultaneously and we would be able to decide which system is the best for which topic, the performance would rise by about 0.05 (absolute, from 0.1958 to 0.2473, relative 26%). The combination of these two assumptions would lead to a MAP of 0.3063.

The main questions are: how to find difficult topics for our system and how to decide, which system resp. which parameters of the retrieval engine fit best to which kind of topic. The answers are as difficult as the prediction of topic developers, which topics are difficult and which are easy resp. easier. It is almost impossible to predict, if a topic is difficult or not [5].

Table 4 presents the most difficult topics for Hildesheim at GeoCLEF 2007 (for German) compared to all systems.

Table 4. Difficult Topics

Topic	Title-DE	Title-EN	MAP HI (DE) (average)	MAP all (DE) (average)	No. of relevant Docs
51	Erdöl- und Gasförderung zwischen dem UK und dem Europäischen Festland	Oil and gas extraction found between the UK and the Continent	0.0099	0.0866	30
52	Verbrechen in der Gegend von St. Andrews	Crime near St Andrews	0.0000	0.0000	0
53	Wissenschaftliche Forschung an Universitäten der schottischen Ostküste	Scientific research at east coast Scottish Universities	0.0082	0.0778	10
57	Whiskyherstellung auf den schottischen Inseln	Whisky making in the Scottish Islands	0.0048	0.0792	1
70	Touristen-Attraktionen in Norditalien	Tourist attractions in Northern Italy	0.0054	0.0639	23

⁴ Worst topics for all systems: 52, 55, 64, 66, 70.

Obviously we can disregard topic 52. The assessment did not find any relevant document for this topic, every system was rated with MAP of 0.

It is very interesting that all these topics show a very small number of relevant assessed documents. In fact a medium positive correlation (+0.5) of the relevant assessed documents and the rank of the topic is proved. Unfortunately, this is not a suitable measure for a system decision whether a topic is difficult or not because the assessed values are of course only available after the assessment.

But problems could be proved at the index of geo-terms. We find the term “Scottish” in topics 53 and 57 in the geo-terms of the index (in 299 documents), but the term “schottisch” (German for “Scottish”) cannot be found in any document. Similar for topic 51: the geo-index shows 5797 documents for the term “Europa”, for “europäisch” (German for “European”) unfortunately none. Adjectives like “europäisch” or “scottish” are obviously difficult to identify as geographic terms for the Named Entity Recognition System and so they are not found in our geo-index. Similar problems about stemming of adjectives have already been identified by Savoy [6].

Considering successful topics (54, 58, 59, 69, 75) on the other hand, we find geo-terms such as “Nordeuropa”, “London”, “Himalaya” and “Birma” in the geo-index very often.

As a consequence, it is necessary to correct the named entity model for geographic terms. If the term “Scottish” is retrieved, geo-terms like “Scotland” have to be included in the geo-index, the stemming of geographical terms could be modified compared to that of other terms.

Another approach is a fusion of the different machines with the different parameters. Considering the separate runs independently of each other, the run HIMODENE2NA performed best 14 times, HIMODENE2 3 times and HIMODENE3 twice and HIMODEBASE five times⁵.

As already mentioned a fusion of the kind, that the result of each run with the best result for a topic was selected resp. got at least a higher weight within the fusion, 5% of absolute improvement at the MAP would be achieved. A possible fusion approach would be the MIMOR (Multiple Indexing and Method-Object Relations, [7]) approach, which has already been tested at former CLEF campaigns [8].

5 Outlook

Optimized Geo-NE models seem to have a positive effect on retrieval quality for monolingual tasks, but it seems to be very difficult to reach significant improvements with only changing the parameters of the BRF. For future experiments, we intend on one hand to integrate geographic ontologies to expand entities with neighbouring places, villages and regions and perform a “geographic stemming” in order to include terms like “Scotland” to the query even if the term “schottisch” is searched. On the other hand, we will reintegrate the fusion approach of MIMOR to merge the result lists of the different systems resp. runs with different parameters.

⁵ As above, Topic 52 has been disregarded, consequently the sum is 24 instead of 25.

References

1. Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xing, X.: GeoCLEF 2007: The CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, Springer, Heidelberg (2008)
2. Bischoff, K., Mandl, T., Womser-Hacker, C.: Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
3. Bischoff, K., Mandl, T., Kölle, R., Womser-Hacker, C.: Geographische Bedingungen im Information Retrieval: Neue Ansätze in Systementwicklung und Evaluierung. In: Oßwald, A., Stempfhuber, M., Wolff, C. (eds.) Open Innovation – neue Perspektiven im Kontext von Information und Wissen? Proc 10. Internationales Symposium für Informationswissenschaft (ISI 2007), Konstanz, 30. Köln Mai - 1. Juni 2007, pp. 15–26. Universitäts verlag [Schriften zur Informationswissenschaft 46] (2007)
4. Gey, F., Larson, R., Sanderson, M., Bishoff, 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: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 852–876. Springer, Heidelberg (2007)
5. Eguchi, K., Kando, N., Kuriyama, K.: Sensitivity of IR Systems Evaluation to Topic Difficulty. In: Araujo, C.P.S., Rodríguez, M.G. (eds.) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, May 29–31, pp. 585–589. ELRA, Paris (2002)
6. Savoy, J.: Why do successful search systems fail for some topics. In: Proceedings of the ACM Symposium on Applied Computing, SAC 2007, Seoul, Korea, March 11 - 15, 2007, pp. 872–877. ACM Press, New York (2007)
7. Womser-Hacker, C.: Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Habilitationsschrift. Universität Regensburg, Informationswissenschaft (1997)
8. Hackl, R., Kölle, R., Mandl, T., Ploedt, A., Scheufen, J.-H., Womser-Hacker, C.: Multilingual Retrieval Experiments with MIMOR at the University of Hildesheim. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 166–173. Springer, Heidelberg (2004)

GIR Experiments with Forostar

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. In this paper we present Forostar, our GIR system. Forostar augments a traditional IR VSM approach with geographic information.

We detail our placename disambiguation and geographic relevance ranking methods, as well as how textual and geographic relevance assessments are combined. The paper concludes with an analysis of our results including significance testing where we show our baseline method, in fact, to be best. Finally we identify weaknesses in our approach and ways in which the system could be optimised and improved.

1 Introduction

This paper describes experiments performed on the GeoCLEF corpus using our GIR system, Forostar [1]. We test the hypothesis that by combining disambiguated locations with textual terms in a single vector space, we can improve over standard IR.

In Section 2 we outline how we index the GeoCLEF corpus and the three field types: Text, Named Entity and Geographic. We then describe how the manually constructed queries are expanded and submitted to the query engine. Section 3 describes and justifies the placename disambiguation methods and geographic relevance ranking methods in more detail. In Section 4 we describe our experiments followed by the results in Section 5. Finally Section 6 analyses the weaknesses of our system and identifies areas for improvement.

2 System

Forostar is our ad-hoc Geographic Information Retrieval system. At indexing time, documents are analysed and named entities are extracted. Named entities tagged as locations are then disambiguated using our co-occurrence model. The free-text fields, named entities and disambiguated locations are then indexed by Lucene. In the querying stage we combine the relevance scores assigned to the Geographic fields and Textual fields using the Vector Space Model. Fields designated as containing more information (i.e. The Headline) have a *boost* value assigned to them.

2.1 Indexing

The indexing stage of Forostar begins by extracting named entities from text using ANNIE, the Information Extraction engine bundled with GATE. GATE is Sheffield University's General Architecture for Text Engineering [2].

Named Entity Fields. We index all the named entities categorised by GATE in a "Named Entity" field in Lucene (e.g. "Police," "City Council," or "President Clinton"). The named entities tagged as Locations by ANNIE, we index as "Named Entity – Location" (e.g. "Los Angeles," "Scotland" or "California") and as a Geographic Location (described later in this section). The body of the GeoCLEF articles and the article titles are indexed as text fields.

Text Fields. Text fields are pre-processed by a customised analyser similar to Lucene's default analyser [3]. Text is split at white space into tokens, the tokens are then converted to lower case, stop words are discarded and the remaining tokens are stemmed with the "Snowball Stemmer". The processed tokens are stored in Lucene's inverted index.

Geographic Fields. The locations tagged by the named entity recogniser are passed to the disambiguation system. We have implemented a simple disambiguation method based on heuristic rules. For each placename being classified

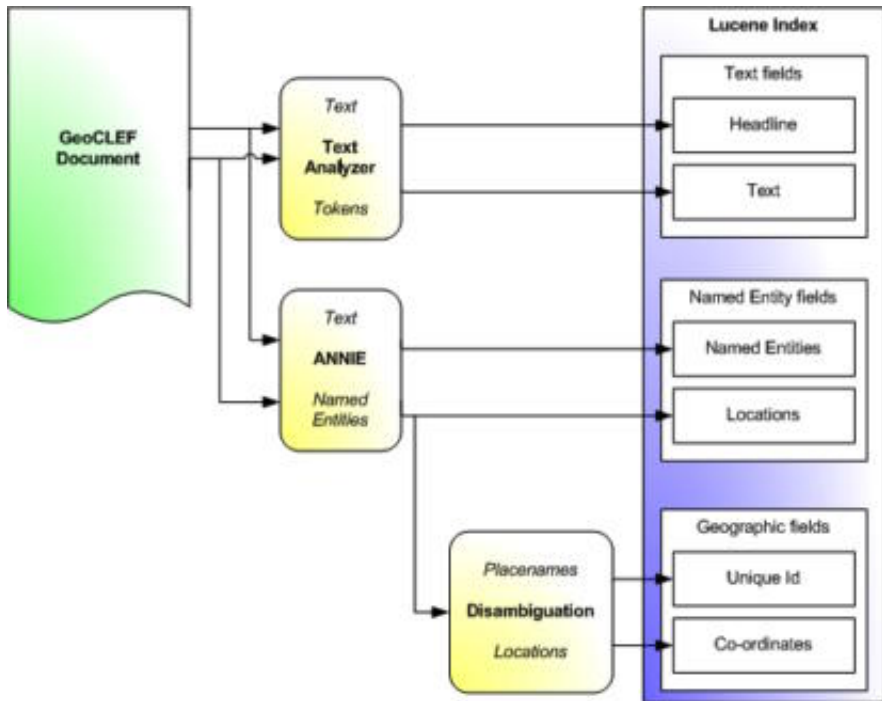


Fig. 1. Building the Lucene Index

we build a list of candidate locations. If there are multiple candidate locations we consider the placename ambiguous. If the placename being classified is followed by a referent location this can often cut down the candidate locations enough to make the placename unambiguous. If the placename is not followed by a referent location or is still ambiguous we disambiguate it as the most commonly occurring location with that name.

Topological relationships between locations are looked up in the Getty Thesaurus of Geographical Names (TGN) [4]. Statistics on how commonly different placenames refer to different locations and a set of synonyms for each location are harvested from our Geographic Co-occurrence model, which in turn is built by crawling Wikipedia [5].

Once placenames have been mapped to unique locations in the TGN, they need to be converted into *Geographic fields* to be stored in Lucene. We store locations in two fields:

- **Coordinates.** The coordinate field is simply the latitude and longitude as read from the TGN.
- **Unique strings.** The unique string is the unique id of this location, preceded with the unique id of all the parent locations, separated with slashes. Thus the unique string for the location “London, UK” is the unique id for London (7011781), preceded by its parent, Greater London (7008136), preceded by its parent, Britain (7002445)... until the root location, the World (1000000) is reached. Giving the unique string for London as 1000000\1000003\7008591\7002445\7008136\7011781.

Note that the text, named entity and geographic fields are **not orthogonal**. This has the effect of multiplying the impact of terms occurring in multiple fields. For example, if the term “London” appears in the text, the token “london” will be indexed in the text field. “London” will be recognised by ANNIE as a Named Entity and tagged as a location (and indexed as Location Entity, “London”). The Location Entity will then be disambiguated as location “7011781” and corresponding geographic fields will be added.

Previous experiments conducted on the GeoCLEF data set in [6] showed improved results from having overlapping fields. We concluded from these experiments that the increased weighting given to locations caused these improvements.

2.2 Querying

The querying stage of Forostar is a two step process. First, manually constructed queries are expanded and converted into Lucene’s bespoke querying language; then we query the Lucene index with these expanded queries and perform blind relevance feedback on the result.

Manually Constructed Query. The queries are manually constructed in a similar structure to the Lucene index. Queries have the following parts: a text field, a Named Entity field and a location field. The text field contains the query with no alteration. The named entity field contains a list of named entities

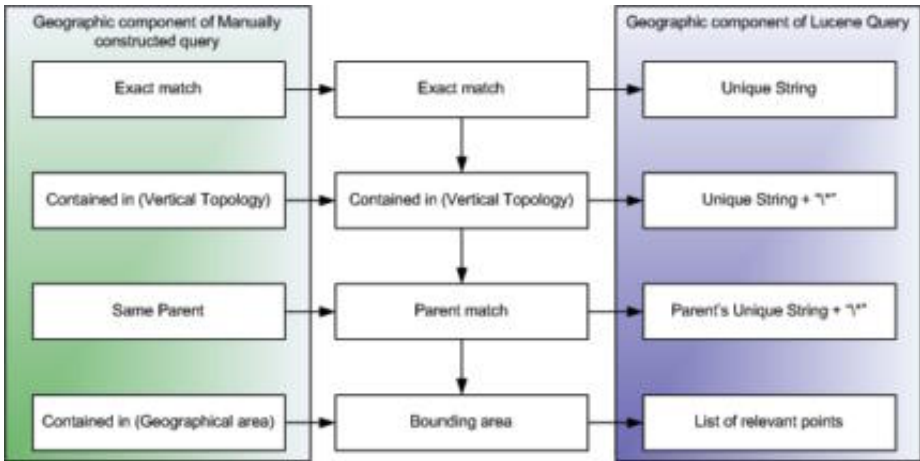


Fig. 2. Expanding the geographic queries

referred to in the query (manually extracted). The location field contains a list of location–relationship pairs. These are the locations contained in the query and their relationship to the location being searched for.

A location can be specified either with a placename (optionally disambiguated with a referent placename), a bounding box, a bounding circle (centre and radius), or a geographic feature type (such as “lake” or “city”). A relationship can either be “exact match,” “contained in (vertical topology),” “contained in (geographic area),” or “same parent (vertical topology)”. The negation of relationships can also be expressed i.e. “excluding,” “outside,” etc.

We believe such a manually constructed query could be automated with relative ease in a fashion similar to the processing that documents go through when indexed. This was not implemented due to time constraints.

Expanding the Geographic Query. The geographic queries are expanded in a pipeline. The location–relation pairs are expanded in turn. The relation governs at which stage the location enters the pipeline. At each stage in the pipeline the geographic query is added to. At the first stage an exact match for this location’s unique string is added: for “London” this would be 1000000\1000003\7008591\7002445\7008136\7011781. Then places within the location are added, this is done using Lucene’s wild-card character notation: for locations in “London” this becomes 1000000\1000003\7008591\7002445\7008136\7011781*. Then places sharing the same parent location are added, again using Lucene’s wild-card character notation. For “London” this becomes all places within “Greater London,” 1000000\1000003\7008591\7002445\7008136*. Finally, the coordinates of all the locations falling *close* to this location are added. A closeness value can be set manually in the location field, however default values are based on feature type (default values were chosen by the authors). The feature listed in the Getty

TGN for “London” is “Administrative Capital,” the default value of closeness for this feature is 100km.

Combining Using the VSM. A Lucene query is built using the text fields, named entity fields and expanded geographic fields. The text field is processed by the same analyzer as at indexing time and compared to both the text and headline fields in the Lucene index. We define a separate boost factor for each field. These boost values were set by the authors during initial iterative tests (they are comparable to similar weighting in past GeoCLEF papers [7,8]). The headline had a boost of 10, the text a boost of 7, named entities a boost of 5, geographic unique string a boost of 5 and geographic coordinates a boost of 3. The geographic, text and named entity relevance are then combined using Lucene’s Vector Space Model.

We perform *blind relevance feedback* on the text fields only. To do this, the whole expanded query is submitted to the Lucene query engine, and the resulting answer set’s top 10 documents considered relevant. The top occurring terms in these documents with more than 5 occurrences are added to the text parts of the query. A maximum of 10 terms are added. The final expanded query is re-submitted to the query engine and our final results are returned.

3 Geographic Retrieval

Forostar allows us to perform experiments on placename disambiguation and geographic relevance ranking. Our geographic model represents locations as points. We choose a point representation over a more accurate polygon representation for several reasons: It makes minimal appreciable difference for queries at the small scale (city or county); Egenhofer and Mark’s *topology matters metrics refine* premise [9] suggests that for queries of a larger scale than city or county topology is of greater importance than distance; and far more point data is available. We represent each location referred to in a document with a single point rather than constructing an encompassing footprint because, we argue, if several locations are referred to in a document, it does not imply locations occurring between the referenced locations are relevant.

3.1 Placename Disambiguation

As discussed in Section 2.1 our placename disambiguation is performed using simple heuristic rules. A key part of the disambiguation is our default gazetteer. The default gazetteer is used to disambiguate placenames that are not immediately followed by a referent placename. The default gazetteer is a many-to-one mapping of Placenames to locations (i.e. for every placename there is a single location). We extract our default gazetteer from a co-occurrence model built from Wikipedia. The generation and analysis of the co-occurrence model is described in [5].

3.2 Geographic Relevance

In Section 2.1 our geographic relevance strategy is described. In this section we provide a justification for the methods used. We have 4 types of geographic relations each expanded differently:

- ‘Exact Match,’ the motivation behind this is that the most relevant documents to a query will mention the location being searched for;
- ‘Contained in (Vertical Topology)’ assumes locations within a location being searched for are relevant, for example ‘London’ will be relevant to queries which search for ‘England’;
- Locations that share the same parent, these locations are topologically close. For example a query for ‘Wales’ would consider ‘Scotland’, ‘England’ and ‘Northern Ireland’ relevant;
- The final method of geographic relevance is defining a viewing area, all locations within a certain radius are considered relevant.

Each geographic relation is considered of greater importance than the following one. This follows Egenhofer and Mark’s *Topology Matters, Metrics Refine* premise. The methods of greater importance are expanded in a pipeline as illustrated in Figure 2. The expanded query is finally combined by Lucene using the Vector Space Model.

4 Experiments

We compared four methods of query construction. All methods query the same index.

- **Standard Text (Text).** This method only used the standard text retrieval part of the system. The motivation for this method was to evaluate our text retrieval engine and provide a baseline.
- **Text with geographic entities removed (TextNoGeo).** For this method we manually removed the geographic entities from the text queries to quantify the importance of ambiguous geographic entities. The results produced by this method should be orthogonal to the results produced by the *Geo* method.
- **Geographic Entities (Geo).** The *Geo* method uses only the geographic entities contained in a query, these are matched ambiguously against the named entity index and unambiguously against the geographic index. Ranking is performed using the geographic relevance methods described in Section 3.2.
- **Text and geographic entities (Text + Geo).** Our combined method combines elements of textual relevance with geographic relevance using the Vector Space Model. It is a combination of the *Text* and *Geo* methods. Our hypothesis is that it will show an improvement over the other tested methods.

Our hypothesis is that a combination of Text and Geographic relevance will give the best results as it uses the most information to discover documents relevant to

the query. The Standard Text method should provide a good baseline to compare this hypothesis against and the orthogonal *Geo* and *TextNoGeo* entries should help us interpret where the majority of the information is held.

5 Results

The experimental results are displayed in Table 1. Surprisingly, the *Text* result is the best, achieving a Mean Average Precision (MAP) of 0.185, with a confidence greater than 99.95% using the Wilcoxon signed rank test [10]. The *Text+Geo* method is better than the *TextNoGeo* method with a confidence greater than 95%. The *Geo* results are the worst with a confidence greater than 99.5%.

Table 1. Mean Average Precision of our four methods

Text 0.185
TextNoGeo 0.099
Geo 0.011
Text+Geo 0.107

74.9% of named entities tagged by ANNIE as locations were mapped to locations in the default gazetteer. This is consistent with the prediction of $\sim 75\%$ made in [5].

Some brief observations of the per query results shows that the *Text+Geo* results are better than *Geo* in all except one case, while the *Text* results are better in all except two cases. The largest variation in results (and smallest significant difference) is the *Text+Geo* and the *TextNoGeo* results.

6 Conclusions

Surprisingly the *Text* method achieved significantly better results than the combination of textual and geographic relevance. We attribute the relatively poor results of the *Text+Geo* method to the way the textual and geographic relevance were combined.

The separate types of geographic relevance and the textual relevance were all combined within Lucene’s vector space model with no normalisation. The motivation behind this was that using Lucene’s term boosting we should be able to give greater weighting to text terms. The difference in information between the *Text+Geo* method and *Text* method are captured in the *Geo* method. Observations of the per query results show that in cases where the *Geo* method performed poorly and the *Text* method performed well, the *Text+Geo* method performed poorly. The intention of combining the two methods was to produce synergy, however, in reality the *Geo* method undermined the *Text* results.

The *Geo* method alone performed poorly compared to the other methods. However, when considering the only information provided in these queries is

geographic information (generally a list of placenames), the results are very promising. The highest per query result achieved by the *geo* method had an average precision of 0.097. Further work is needed to evaluate the accuracy of the placename disambiguation. Currently we have only quantified that 74.9% of locations recognised by ANNIE are disambiguated. We have not yet evaluated the disambiguation accuracy or the proportion of locations that are missed by ANNIE.

In future work we would like to repeat the combination experiment detailed in this paper, however separating the geographic relevance and textual relevance into two separate indexes. Similarity values with respect to a query could be calculated for both indexes, normalised and combined in a weighted sum. A similar approach to this was taken in GeoCLEF 2006 by Martins et al. [7].

References

1. Mandl, T., Gey, F., Nunzio, G.D., Ferro, N., Larson, R., Sanderson, M., Santos, D., Womser-Hacker, C., Xie, X.: GeoCLEF 2007: the CLEF 2007 cross-language geographic information retrieval track overview. In: Working Notes for the CLEF Workshop (2007)
2. Cunningham, H., Maynard, D., Tablan, V., Ursu, C., Bontcheva, K.: Developing language processing components with GATE. Technical report, University of Sheffield (2001)
3. Apache Lucene Project (2007) (Accessed 1 August 2007), <http://lucene.apache.org/java/docs/>
4. Harping, P.: User's Guide to the TGN Data Releases. The Getty Vocabulary Program, 2nd edn (2000)
5. Overell, S., Rüger, S.: Geographic co-occurrence as a tool for GIR. In: CIKM Workshop on Geographic Information Retrieval (2007)
6. Overell, S., Magalhães, J., Rüger, S.: Forostar: A system for GIR. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 930–937. Springer, Heidelberg (2007)
7. Martins, B., Cardoso, N., Chaves, M., Andrade, L., Silva, M.: The University of Lisbon at GeoCLEF 2006. In: Working Notes for the CLEF Workshop (2006)
8. Ruiz, M., Shapiro, S., Abbas, J., Southwick, S., Mark, D.: UB at GeoCLEF 2006. In: Working Notes for the CLEF Workshop (2006)
9. Egenhofer, M., Mark, D.: Naive geography. In: The 1st Conference on Spatial Theory (COSIT) (1995)
10. Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Annual international ACM SIGIR Conference, pp. 329–338 (1993)

Morpho Challenge Evaluation Using a Linguistic Gold Standard

Mikko Kurimo, Mathias Creutz, and Matti Varjokallio

Adaptive Informatics Research Centre, Helsinki University of Technology,
P.O.Box 5400, FIN-02015 TKK, Finland
{mikko.kurimo,mathias.creutz,matti.varjokallio}@tkk.fi
<http://www.cis.hut.fi/morphochallenge2007/>

Abstract. In Morpho Challenge 2007, the objective was to design statistical machine learning algorithms that discover which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, information retrieval, and statistical language modeling. Because in unsupervised morpheme analysis the morphemes can have arbitrary names, the analyses are here evaluated by a comparison to a linguistic gold standard by matching the morpheme-sharing word pairs. The data sets were provided for four languages: Finnish, German, English, and Turkish and the participants were encouraged to apply their algorithm to all of them. The results show significant variance between the methods and languages, but the best methods seem to be useful in all tested languages and match quite well with the linguistic analysis.

Keywords: Morphological analysis, Machine learning.

1 Introduction

The scientific objectives of the Morpho Challenge 2007 were: to learn of the phenomena underlying word construction in natural languages, to advance machine learning methodology, and to discover approaches suitable for a wide range of languages. The suitability for a wide range of languages is becoming increasingly important, because language technology methods need to be quickly and as automatically as possible extended to new languages that have limited previous resources. That is why learning the morpheme analysis directly from large text corpora using unsupervised machine learning algorithms is such an attractive approach and a very relevant research topic today.

The Morpho Challenge 2007 is a follow-up to Morpho Challenge 2005 (Unsupervised Segmentation of Words into Morphemes) [1]. In the Morpho Challenge 2005 the task was to design an unsupervised statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. In addition to comparing the obtained morphemes to a linguistic gold standard, their usefulness was evaluated by using them for training statistical language models that were tested in speech recognition.

In the Morpho Challenge 2007 the focus was more general in not only segmenting words, but also to perform *morpheme analysis* of the word forms. For instance, the English words "boot, boots, foot, feet" might obtain the analyses "boot, boot + plural, foot, foot + plural", respectively. In linguistics, the concept of morpheme does not necessarily directly correspond to a particular word segment but to an abstract class. For some languages there exist carefully constructed linguistic tools for this kind of analysis, although not for many, but statistical machine learning methods may still discover interesting alternatives that may rival even the most careful linguistically designed morphologies.

The problem of learning the morphemes directly from large text corpora using an unsupervised machine learning algorithm is clearly a difficult one. First the words should be somehow segmented into meaningful parts, and then these parts should be clustered in the abstract classes of morphemes that would be useful for modeling. It is also challenging to learn to generalize the analysis to rare words, because even the largest text corpora are very sparse, a significant portion of the words may occur only once. Many important words, for example proper names and their inflections or some forms of long compound words, may also not exist in the training material at all, and their analysis is often even more challenging. Benefits for successful morpheme analysis, in addition to obtaining a set of basic vocabulary units for modeling, can be seen for many important tasks in language technology. The additional information included in the units can provide support for building more sophisticated language models, for example, in speech recognition [2], machine translation [3], and IR [4].

Arranging a meaningful evaluation of the unsupervised morpheme analysis algorithms is not straight-forward, because in unsupervised morpheme analysis the labels of the morphemes are arbitrary and not likely to directly correspond to the linguistic morpheme definitions. In this challenge we performed two complementary evaluations, one including a comparison to linguistic morpheme analysis, and another in a practical application where morpheme analysis might be used. In the first evaluation, described in this paper, the proposed morpheme analyses were compared to a linguistic gold standard citecreutz04.tr by counting the matching morpheme-sharing word pairs. In this way we did not have to try to match the labels of the morphemes directly, but only to measure if the proposed algorithm can find the correct word pairs that share common morphemes. In the second evaluation, described in a companion paper [5], IR experiments were performed using the CLEF resources, but the words in the documents and queries replaced by their proposed morpheme representations and the search based on morphemes instead of words.

2 Task and Data Sets

The task for the participants was set to return the unsupervised morpheme analysis of every word form contained in a long word list supplied by the organizers for each test language. The participants were pointed to corpora in which the words occur, so that their algorithms may utilize information about word

context. Data sets and evaluations were provided for the same three languages as in the Morpho Challenge 2005: Finnish, English, and Turkish, plus one new language, German. To achieve the goal of designing language independent methods, the participants were encouraged to submit results in all these languages.

The participants were allowed to supply several interpretations of each word, because many words do have them: e.g., the word "flies" may be the plural form of the noun "fly" (insect) or the third person singular present tense form of the verb "to fly". Thus the analysis could be as: "FLY_N +PL, FLY_V +3SG". The existence of alternative analyses made the task challenging, and it was left to the participants to decide how much effort they put into this aspect of the task.

The English and German gold standards were based on the CELEX data base¹. The Finnish gold standard was based on the two-level morphology analyzer FINTWOL from Lingsoft², Inc. The Turkish gold-standard analyses was obtained from a morphological parser developed at Bogazici University³ [6,7]; it is based on Oflazer's finite-state machines, with a number of changes. Examples of the gold standard analysis are shown in Table 2.

To encourage for further and more detailed evaluations using the gold standard analysis, the gold standard labels that correspond to affixes were distinguished from stems by marking them with an initial plus sign (e.g., +PL, +PAST). The stem were labeled by an intuitive string, usually followed by an underscore character (_) and a part-of-speech tag, e.g., "baby_N", "sit_V". In many cases, especially in English, the same morpheme can function as different parts-of-speech; e.g., the English word "force" can be a noun or a verb. However, there was not really a need for the participant's algorithm to distinguish between different meanings or syntactic roles of the discovered stem morphemes.

3 Participants and the Submissions

6 research groups submitted the segmentation results obtained by 12 different algorithms and 8 of them participated in all four test languages. All the submitted algorithms are listed in Table 1. In general, the submissions were all interesting and all of them met the exact specifications given and were able to get properly evaluated. Additionally, we evaluated a public baseline method called "Morfessor Categories-MAP" (or here just "Morfessor MAP" or "Morfessor", for short) developed by the organizers [8]. Naturally, the Morfessors competed outside the main competition and the results were included only as a reference.

Table 2 shows an example analysis and some statistics of each submission including the average amount of alternative analysis per word, the average amount of morphemes per analysis, and the total amount of morpheme types. The total amount of word types were 2,206,719 (Finnish), 617,298 (Turkish), 1,266,159 (German), and 384,903 (English). The Turkish word list was extracted in 1 million sentences, but the other lists from 3 million sentences per each language. In

¹ <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC96L14>

² <http://www.lingsoft.fi/>

³ http://www.boun.edu.tr/index_eng.html

Table 1. The submitted algorithms and reference methods

Algorithm	Authors	Affiliation
“Bernhard 1, 2”	Delphine Bernhard	TIMC-IMAG, F
“Bordag 5, 5a”	Stefan Bordag	Univ. Leipzig, D
“McNamee 3, 4, 5”	Paul McNamee and James Mayfield	JHU, USA
“Zeman ”	Daniel Zeman	Karlova Univ., CZ
“Monson Morfessor”	Christian Monson et al.	CMU, USA
“Monson ParaMor”	Christian Monson et al.	CMU, USA
“Monson ParaMor-M”	Christian Monson et al.	CMU, USA
“Pitler”	Emily Pitler and Samarth Keshava	Univ. Yale, USA
“Morfessor MAP”	The organizers	Helsinki Univ. Tech, FI
“Tepper”	Michael Tepper	Univ. Washington, USA
“Gold Standard”	The organizers	Helsinki Univ. Tech, FI

these word lists, the gold standard analysis were available for 650,169 (Finnish), 214,818 (Turkish), 125,641 (German), and 63,225 (English) words.

The algorithms by Bernhard, Bordag and Pitler were the same or improved versions from the previous Morpho Challenge [1]. Monson and Zeman were new participants who also provided several alternative analysis for most words. The most different approach was McNamee’s algorithm which did not aim at morpheme analysis, but mainly to find a representative substring for each word type. In Table 2, the statistics are only shown for “McNamee 3”. Noteworthy in Table 2 is also that the size of the morpheme lexicon varied a lot in different algorithms.

4 New Evaluation Method

For each language, the morpheme analyses proposed by the participants’ algorithm were compared against the linguistic gold standard. Since the task at hand involved unsupervised learning, it could not be expected that the algorithm came up with morpheme labels that exactly correspond to the ones designed by linguists. That is, no direct comparison took place between labels as such (the labels in the proposed analyses vs. labels in the gold standard). What could be expected, however, was that two word forms that contained the same morpheme according to the participants’ algorithm also had a morpheme in common according to the gold standard. For instance, in the English gold standard, the words “foot” and “feet” both contain the morpheme “foot_N”. Thus, the goal was that the participants’ algorithm discovered a morpheme that occurred in both these word forms (be it called “FOOT”, “morpheme784”, “foot” or something else).

In practice, the evaluation took place by randomly sampling a large number of word pairs, such that both words in the pair had at least one morpheme in common. The exact constitution of this set of word pairs was not revealed to the participants. In the evaluation, word frequency played no role. Thus, all word pairs were equally important, whether they were frequent or rare. The size of the randomly chosen set of word pairs set varied depending on the size of the

Table 2. Statistics and example morpheme analyses. #anal is the average amount of analysis per word (separated by a comma), #mor the average amount of morphemes per analysis (separated by a space), and lexicon the total amount of morpheme types.

Finnish	Example word: linuxiin	#anal	#mor	lexicon
Bernhard 1	linux_B iin_S	1	3.16	87590
Bernhard 2	linux_B i_S in_S	1	3.89	87915
Bordag 5	linuxiin	1	2.84	517091
Bordag 5a	linuxia.linuxiin	1	2.84	514670
McNamee 3,4,5	xii, uxii, nuxii	1	1	20063
Zeman	linuxiin, linuxii n, linuxi in, linux iin	3.62	1.81	5434453
Morfessor MAP	linux +iin	1	2.94	217001
Gold Standard	linux_N +ILL	1.16	3.29	33754
Turkish	Example word: popUlerliGini	#anal	#mor	lexicon
Bernhard 1	popUler_B liGini_S	1	2.48	86490
Bernhard 2	popUler_B liGini_S	1	2.73	87637
Bordag 5	popUlerliGini	1	2.24	219488
Bordag 5a	popUlerliGini	1	2.24	219864
McNamee 3,4,5	opU, pUle, Ulerl	1	1	19389
Zeman	popUlerliGin i, popUlerliGi ni	3.24	1.76	1205970
Morfessor MAP	pop +U +ler +liGini	1	2.64	114834
Tepper	popU lEr lWK W W	1	2.81	110682
Gold Standard	popUler +DER_lHg +POS2S +ACC, ...	1.99	3.36	21163
German	Example word: zurueckzubehalten	#anal	#mor	lexicon
Bernhard 1	zurueckzu_P behalt_B en_S	1	3.12	56173
Bernhard 2	zurueckzu_P behalt_B e_S n_S	1	3.72	53497
Bordag 5	zu rueck zu be halt en	1	2.99	267680
Bordag 5a	zu rueck zu be ehalt.hale.halt.halte.helt en	1	2.99	266924
McNamee 3,4,5	kzu, kzub, kzube	1	1	16633
Zeman	zurueckzubehalten, zurueckzubehalte n, ...	4.15	1.81	4094228
Monson Paramor-M	+zurueck/P +zu/P +be/P halten/B, ...	2.91	2.20	1191842
Monson ParaMor	zurueckzub +ehalten, zurueckzube +halten	1.91	1.72	1001441
Monson Morfessor	+zurueck/P +zu/P +be/P halten/B	1	3.10	166963
Morfessor MAP	zurueck zu be halten	1	3.06	172907
Gold Standard	zurueck_B zu be halt_V +INF	1.30	2.97	14298
English	Example word: baby-sitters	#anal	#mor	lexicon
Bernhard 1	baby_P -L sitt_B er_S s_S	1	2.61	55490
Bernhard 2	baby_P -L sitt_B er_S s_S	1	2.90	52582
Bordag 5	baby sitters	1	1.97	190094
Bordag 5a	baby sitters	1	1.97	189568
McNamee 3,4,5	by-, aby-, y-sit	1	1	15212
Zeman	baby-sitter s, baby-sitt ers	3.18	1.74	905251
Monson Paramor-M	+baby-/P sitter/B +s/S, bab+y, ...	3.42	1.93	386257
Monson ParaMor	bab+y, sit+ters, sitt+ers, sitte+rs, sitter+s	2.42	1.88	233981
Monson Morfessor	+baby-/P sitter/B +s/S	1	2.07	137973
Pitler	baby- sitt ers	1	1.57	211475
Morfessor MAP	baby - sitters	1	2.12	132086
Tepper	baby - sit ers	1	2.53	99937
Gold Standard	baby_N sit_V er_s +PL	1.10	2.13	16902

word lists and gold standard given in the previous section: 200,000 (Finnish), 50,000 (Turkish), 50,000 (German), and 10,000 (English) word pairs.

As the evaluation measure, we applied *F-measure*, which is the harmonic mean of *Precision* and *Recall*: $F\text{-measure} = 1/(1/Precision + 1/Recall)$.

Precision was here calculated as follows: A number of word forms were randomly sampled from the result file provided by the participants; for each morpheme in these words, another word containing the same morpheme were chosen from the result file by random (if such a word existed). We thus obtained a number of word pairs such that in each pair at least one morpheme was shared between the words in the pair. These pairs were compared to the gold standard; a point was given for each word pair that really had a morpheme in common according to the gold standard. The total number of points was then divided by the total number of word pairs.

Recall was calculated analogously to precision: A number of word forms were randomly sampled from the gold standard file; for each morpheme in these words, another word containing the same morpheme were chosen from the gold standard by random (if such a word existed). The word pairs were then compared to the analyses provided by the participants; a point was given for each sampled word pair that had a morpheme in common also in the analyses proposed by the participants' algorithm. The total number of points was then divided by the total number of sampled word pairs.

For words that had several alternative analyses, as well as for word pairs that had more than one morpheme in common, the normalization of the points was carried out in order not to give these words considerably more weight in the evaluation than "less complex" words. The words were normalized by the number of alternative analyses and the word pairs by the number of matching morphemes. The evaluation script⁴ was provided for the participants to check their morpheme analysis against the available gold standard samples.

5 Results and Discussions

The precision, recall and F-measure percentages obtained in the evaluation for all the test languages are shown in Table 3. Two reference results are given below each table. *Morfessor Categories-Map*: The same Morfessor Categories-Map as described in Morpho Challenge 2005 [9] was used for the unsupervised morpheme analysis. Each morpheme was also automatically labeled as prefix, stem or suffix by the algorithm. *Tepper*: A hybrid method developed by Michael Tepper [10] was utilized to improve the morpheme analysis reference obtained by our Morfessor Categories-MAP.

For the Finnish task the winner (measured by F-measure) was the algorithm "Bernhard 2". It did not reach a particularly high precision, but the recall and the F-measure were clearly superior. It was also the only algorithm that won the "Morfessor MAP" reference. For the Turkish task the competition was much

⁴ Available at <http://www.cis.hut.fi/morphochallenge2007/>

Table 3. The submitted unsupervised morpheme analysis compared to gold standard

Finnish	PRECISION	RECALL	F-MEASURE
Bernhard 2	59.65%	40.44%	48.20%
Bernhard 1	75.99%	25.01%	37.63%
Bordag 5a	71.32%	24.40%	36.36%
Bordag 5	71.72%	23.61%	35.52%
Zeman	58.84%	20.92%	30.87%
McNamee 3	45.53%	8.56%	14.41%
Morfessor MAP	76.83%	27.54%	40.55%
Turkish	PRECISION	RECALL	F-MEASURE
Zeman -	65.81%	18.79%	29.23%
Bordag 5a	81.31%	17.58%	28.91%
Bordag 5	81.44%	17.45%	28.75%
Bernhard 2	73.69%	14.80%	24.65%
Bernhard 1	78.22%	10.93%	19.18%
McNamee 3	65.00%	10.83%	18.57%
Morfessor MAP	76.36%	24.50%	37.10%
Tepper	70.34%	42.95%	53.34%
German	PRECISION	RECALL	F-MEASURE
Monson Paramor-Morfessor	51.45%	55.55%	53.42%
Bernhard 2	49.08%	57.35%	52.89%
Bordag 5a	60.45%	41.57%	49.27%
Bordag 5	60.71%	40.58%	48.64%
Monson Morfessor	67.16%	36.83%	47.57%
Bernhard 1	63.20%	37.69%	47.22%
Monson ParaMor	59.05%	32.81%	42.19%
Zeman -	52.79%	28.46%	36.98%
McNamee 3	45.78%	9.28%	15.43%
Morfessor MAP	67.56%	36.92%	47.75%
English	PRECISION	RECALL	F-MEASURE
Bernhard 2	61.63%	60.01%	60.81%
Bernhard 1	72.05%	52.47%	60.72%
Pitler	74.73%	40.62%	52.63%
Monson Paramor-Morfessor	41.58%	65.08%	50.74%
Monson ParaMor	48.46%	52.95%	50.61%
Monson Morfessor	77.22%	33.95%	47.16%
Zeman	52.98%	42.07%	46.90%
Bordag 5a	59.69%	32.12%	41.77%
Bordag 5	59.80%	31.50%	41.27%
McNamee 3	43.47%	17.55%	25.01%
Morfessor MAP	82.17%	33.08%	47.17%
Tepper	69.23%	52.60%	59.78%

tighter. The winner was “Zeman”, but Bordag’s both algorithms were very close. The “Morfessor MAP” and “Tepper” reference methods was clearly better than any of the competitors, but all the algorithms (except “Tepper”) seem to have had problems with the Turkish task, because the scores were lower than for other languages. This is interesting, because in the morpheme segmentation task of the previous Morpho Challenge [1] the corresponding Turkish task was not more difficult than the others. The “Monson Paramor-Morfessor” algorithm reached the highest score in the German task, but the “Bernhard 2” who again had the highest recall as in Finnish was quite close. For English, Bernhard’s both algorithms were the clear winners, but also “Pitler” and Monson’s algorithms were able to beat the “Morfessor MAP”.

The significance of the differences in F-measure was analyzed for all algorithm pairs in all evaluations. The analysis was performed by splitting the data into several partitions and computing the results for each partition separately. The statistical significance of the differences between the participants’ algorithms was computed by the Wilcoxon’s Signed-Rank test for comparison of the results in the independent partitions. The results show that almost all differences were statistical significant, only the following pairs were not: Turkish: “Zeman”-“Bordag 5a”, “Bordag 5a”-“Bordag”; German: “Monson Morfessor”-“Bernhard 1”; English: “Bernhard 2”-“Bernhard 1”, “Monson Paramor-Morfessor”-“Monson ParaMor”, “Monson Morfessor”-“Zeman”, “Bordag 5a”-“Bordag”. This result was not surprising since the random word pair samples were quite large and all these result pairs that were not significantly different gave very similar F-measures (less than 0.5 percentage units away). From McNamee’s algorithms only “McNamee 3” is shown in Table 3. These algorithms did not aim at morpheme analysis, but mainly to find a representative substring for each word type that would be likely to perform well in the IR evaluation [5].

The future work in unsupervised morpheme analysis should develop further the clustering of contextually similar units for morphemes that would match better with the grammatical morphemes and thus, improve the recall. Most of the submitted algorithms probably did not take the provided possibility to utilize the sentence context for analyzing the words and finding the morphemes. Although this may not be as important for success in IR than improving the precision, it may provide useful additional information for some keywords.

6 Conclusions and Acknowledgments

The objective of Morpho Challenge 2007 was to design a statistical machine learning algorithm that discovers which morphemes (smallest individually meaningful units of language) words consist of. Ideally, these are basic vocabulary units suitable for different tasks, such as text understanding, machine translation, IR, and statistical language modeling. The current challenge was a successful follow-up to our previous Morpho Challenge 2005, but this time the task was more general in that instead of looking for an explicit segmentation of words, the focus was in the morpheme analysis of the word forms in the data.

The scientific goals of this challenge were to learn of the phenomena underlying word construction in natural languages, to discover approaches suitable for a wide range of languages and to advance machine learning methodology. The analysis and evaluation of the submitted machine learning algorithm for unsupervised morpheme analysis showed that these goals were quite nicely met. There were several novel unsupervised methods that achieved good results in several test languages, both with respect to finding meaningful morphemes and useful units for IR. The algorithms and results were presented in Morpho Challenge Workshop, arranged in connection with CLEF 2007, September 19-21, 2007.

Morpho Challenge 2007 was part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF. We are most grateful to the University of Leipzig for making the training data resources available to the Challenge, and in particular we thank Stefan Bordag for his kind assistance. We are indebted also to Ebru Arisoy for making the Turkish gold standard available to us. Our work was supported by the Academy of Finland in the projects *Adaptive Informatics* and *New adaptive and learning methods in speech recognition* and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

1. Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E., Saraclar, M.: Unsupervised segmentation of words into morphemes - Challenge 2005, an introduction and evaluation report. In: PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes, Venice, Italy (2006)
2. Bilmes, J.A., Kirchhoff, K.: Factored language models and generalized parallel backoff. In: Proceedings of HLT-NAACL, Edmonton, Canada, pp. 4-6 (2003)
3. Lee, Y.S.: Morphological analysis for statistical machine translation. In: Proceedings of HLT-NAACL, Boston, MA, USA (2004)
4. Ziemann, Y., Bleich, H.: Conceptual mapping of user's queries to medical subject headings. In: Proceedings of the 1997 American Medical Informatics Association (AMIA) Annual Fall Symposium (October 1997)
5. Kurimo, M., Creutz, M., Turunen, V.: Morpho Challenge evaluation by IR experiments. In: Peters, C., et al. (eds.) CLEF 2007 Workshop. LNCS, vol. 5152. Springer, Heidelberg (2008)
6. Cetinoglu, O.: Prolog based natural language processing infrastructure for Turkish. M.Sc. thesis, Bogazici University, Istanbul, Turkey (2000)
7. Dutagaci, H.: Statistical language models for large vocabulary continuous speech recognition of Turkish. M.Sc. thesis, Bogazici University, Istanbul, Turkey (2002)
8. Creutz, M., Lagus, K.: Inducing the morphological lexicon of a natural language from unannotated text. In: Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005), Espoo, Finland, pp. 106-113 (2005)
9. Creutz, M., Lagus, K.: Morfessor in the Morpho Challenge. In: PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes, Venice, Italy (2006)
10. Tepper, M.: A Hybrid Approach to the Induction of Underlying Morphology. PhD thesis, University of Washington (2007)

Simple Morpheme Labelling in Unsupervised Morpheme Analysis

Delphine Bernhard

Ubiquitous Knowledge Processing Lab
Computer Science Department
Technische Universität Darmstadt, Germany
`delphine@tk.informatik.tu-darmstadt.de`

Abstract. This paper describes a system for unsupervised morpheme analysis and the results it obtained at Morpho Challenge 2007. The system takes a plain list of words as input and returns a list of labelled morphemic segments for each word. Morphemic segments are obtained by an unsupervised learning process which can directly be applied to different natural languages. Results obtained at competition 1 (evaluation of the morpheme analyses) are better in English, Finnish and German than in Turkish. For information retrieval (competition 2), the best results are obtained when indexing is performed using Okapi (BM25) weighting for all morphemes minus those belonging to an automatic stop list made of the most common morphemes.

1 Introduction

The goal of Morpho Challenge 2007 [1,2] was to develop algorithms able to perform unsupervised morpheme analysis, which consists in automatically discovering a word's morphemes using only minimal resources made up of a list of words and a text corpus in each language. Morphemic segments have to be identified but also labelled and hence disambiguated. On the one hand, morphemes are abstract units which may be realised by several surface forms, i.e. allomorphs. On the other hand, a single surface form may correspond to several homographic morphemes. In order to perform morpheme analysis and correctly label morphemic segments, it is therefore necessary to achieve a mapping between morpheme labels and their surface realisations.

As it will be evidenced later, the system presented in this paper does not solve cases of allomorphy since different surface forms will always be considered as different morphemes. It only partly aims at resolving cases of homography since identified morphemic segments are labelled with one of the following categories: stem/base, prefix, suffix and linking element. The system nevertheless achieved decent results in competition 1 for English, Finnish and German, but results in Turkish were less satisfactory.

The rest of this paper is organised as follows. The algorithm is described in Section 2. Results obtained for competitions 1 and 2 are presented in Section 3. Then, particular assumptions made in the system and pertaining to morpheme

labelling are discussed in Section 4. Finally perspectives for the evolution of the system are given in Section 5.

2 Overview of the Method

The algorithm is mostly identical to the one already presented at Morpho Challenge 2005, apart from a few minor changes. A previous and more detailed description of the system can be found in [3]. The method takes as input a plain list of words without additional information. The output is a labelled segmentation of the input words. Labels belong to one of the following categories: stem, prefix, suffix and linking element. The algorithm can be subdivided into 4 main steps, plus an additional step which may be performed to analyse another data set, given the list of segments learned after step 4.

2.1 Step 1: Extraction of Prefixes and Suffixes

The objective of step 1 is to acquire a list of prefixes and suffixes. The longest words in the input word list are segmented based on the notion of segment predictability. This idea is recurrent in research on the segmentation of words into morphemes (see for instance [4,5,6]). It posits that a morpheme boundary can be hypothesised if it is difficult to predict the character (or string of characters) which follows, knowing an initial string of characters. In the system, segment predictability is modelled by computing the average maximum transition probabilities between all the substrings of a word coalescing at a given position k within the word. The variations of this measure make it possible to identify morpheme boundaries at positions where the average maximum transition probabilities reach a well-marked minimum. Figure 1 depicts the variations of the average maximum transition probabilities for the English word “hyperventilating”. Two morpheme boundaries are identified in this word, which corresponds to the following segmentation: “hyper + ventilat + ing”.

Once a word has been segmented in this fashion, the longest and less frequent amongst the proposed segments is identified as a stem, if this segment also appears at least twice in the word list and at least once at the beginning of a word. In the example of Fig. 1, the segment ‘ventilat’ will be identified as a valid stem.

The identified stem is then used to acquire affixes. All the substrings preceding this stem in the input word list are added to the list of prefixes unless they are longer and less frequent than the stem. Correspondingly, all the substrings following this stem in the word list are added to the list of suffixes unless they are longer and less frequent than the stem. Moreover, one character-long prefixes are eliminated because these often lead to erroneous segmentations in later stages of the algorithm.

This procedure is applied to the longest words in the input word lists. The process of affix acquisition ends when for N running words the number of new affixes among the affixes learned is inferior to the number of affixes which already belong to the list of prefixes and suffixes.

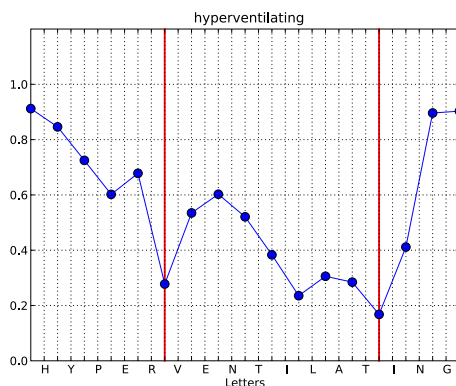


Fig. 1. Variations of the average maximum transition probabilities between substrings of the word “hyperventilating”. Boundaries are marked with a bold vertical line.

2.2 Step 2: Acquisition of Stems

The aim of the second step is to acquire a list of stems, using the prefixes and suffixes which have been previously identified. Stems are obtained by stripping off from each word in the input word list all the possible combinations of prefixes, suffixes and the empty string. In order to reduce the noise induced by such a simple method, some constraints are applied, especially a minimum length threshold of 3 characters for the stem. Note that the stems acquired by this method are not all minimal and may still contain affixes.

2.3 Step 3: Segmentation of Words

In a third step, all the words are segmented. Word segmentation is performed by comparing words which contain the same stem to one another. This consists in finding limits between shared and different segments and results in a segmentation of the words being compared. The outcome can be represented as an alignment graph for each stem. Figure 2 depicts the alignment graph obtained for the words containing the English stem ‘integrat’.

Segments are subsequently labelled with one of the three non-stem types (prefix, suffix, linking element) according to their positions within the word, relatively to the stem. As a result of the alignment, prefixes and suffixes which do not belong to the list of affixes acquired after step 1 may be discovered. A validation procedure, similar to the one proposed by [4], is therefore applied. It consists in checking that the proportion of new affixes in the alignment graph does not exceed some threshold¹. All the segmentations made up of valid morphemic segments are stored.

¹ There are actually two different thresholds, a and b . For details about these thresholds, see [3].

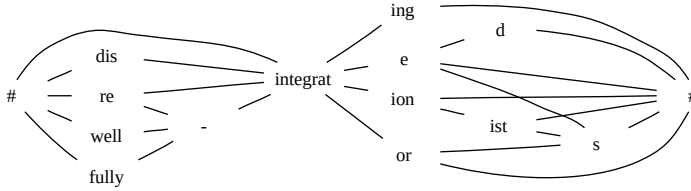


Fig. 2. Example segmentation of words sharing the stem “integrat”

2.4 Step 4: Selection of the Best Segmentation

As a result of Step 3, several different segmentations may have been discovered for each word, since a word may contain more than one potential stem. In order to select the best possible segments, a best-first search strategy is applied on the potential segments of a word, privileging the most frequent segment when given a choice. Some frequency and morphotactic constraints are also checked (e.g. a prefix cannot be directly followed by a suffix, there has to be at least one stem in the proposed segmentation, etc.).

2.5 Step 5: Application of the Learned Segments to a New Data Set (Optional)

The morphemic segments identified after Step 4 can be used to segment any list of words or the whole list of words when segments are learned using only a subset of the list. An A*-like algorithm is used to find the best segmentation for each word, i.e. the segmentation with the lowest global cost which also complies to morphotactic constraints similar to those used at step 4. The global cost for a segmentation is the sum of the costs associated with each segment s_i . Two different segment cost functions have been used resulting in two different submissions for each language:

$$\text{cost}_1(s_i) = -\log \frac{f(s_i)}{\sum_i f(s_i)} \quad (1)$$

$$\text{cost}_2(s_i) = -\log \frac{f(s_i)}{\max_i [f(s_i)]} \quad (2)$$

where $f(s_i)$ is the frequency of s_i .

3 Morpho Challenge 2007 Experiments and Results

The method has been applied to all of the four test languages of Morpho Challenge 2007. Morphemic segments have been learned using only a subset of the word lists provided for competition 1 (the 300,000 most frequent words), without taking into account contextual information found in the text corpora. These

morphemic segments have then been used to segment all the words in the data sets provided both for competition 1 and 2, using either cost_1 or cost_2 .

Moreover, no fine tuning of the three different parameters of the system (N , a and b) has been attempted. Earlier experiments have shown that default values $N=5$, $a=0.8$ and $b=0.1$ are globally reasonable and these have consequently been used for all four languages.

3.1 Results for Competition 1: Morpheme Analysis

In competition 1, the system's analyses have been compared to a linguistic gold standard in Finnish, Turkish, German and English [1]. Table 1 details the precision, recall and F-measure obtained by the system. Method 1 corresponds to results obtained using cost_1 and method 2 to results obtained using cost_2 .

Table 1. Precision %, recall % and F-measure % obtained for competition 1

Language	Method 1			Method 2		
	Precision	Recall	F-measure	Precision	Recall	F-measure
English	72.05	52.47	60.72	61.63	60.01	60.81
Finnish	75.99	25.01	37.63	59.65	40.44	48.20
German	63.20	37.69	47.22	49.08	57.35	52.89
Turkish	78.22	10.93	19.18	73.69	14.80	24.65

As it had already pointed out at Morpho Challenge 2005, results for method 1, using cost_1 , indicate higher precision but lower recall on all datasets. On the whole, better F-measures are obtained with method 2 for all languages. Results in Turkish are well under those obtained in the other languages and are characterised by very poor recall. The recall of most of the other systems which also took part in Morpho Challenge 2007 is below 20% as well for this particular language. One possible explanation for this is that there are more different analyses per word in Turkish (at least in the provided gold standard sample), and therefore more ambiguities have to be solved in the proposed morpheme analyses than in the other languages.

3.2 Results for Competition 2: Information Retrieval

In competition 2, the morphological analyses were used to perform information retrieval experiments. The experimental set-up is described in detail in [2]. Table 2 lists the results obtained for the information retrieval task.

The results obtained by the algorithm strongly depend on the weighting scheme used and are better with the Okapi BM25 weighting and a stop list, whatever the method and the word list used. Moreover, while method 1 performs slightly better than method 2 with the tf-idf weighting, this tendency is reversed with the Okapi weighting (except for German). It is not clear how this could be accounted for but a possible explanation is that method 2, which is less precise, benefits more than method 1 from the removal of the most frequent morphemes.

Table 2. Average precision obtained for competition 2

	English		Finnish		German	
	tf-idf	Okapi	tf-idf	Okapi	tf-idf	Okapi
method 1 - without new	0.2781	0.3881	0.4016	0.4183	0.3777	0.4611
method 1 - with new	0.2777	0.3900	0.3896	0.4681	0.3720	0.4729
method 2 - without new	0.2673	0.3922	0.3984	0.4425	0.3731	0.4676
method 2 - with new	0.2682	0.3943	0.3811	0.4915	0.3703	0.4625

4 Analysis and Discussion

As mentioned in the introduction, the main objective of Morpho Challenge 2007 is to obtain a morpheme analysis of the word forms, which is a lot more demanding than just segmenting words into morphs. A morpheme analysis for a word corresponds to a list of labelled morphemes. A minimal morpheme analysis may consist of a list of unlabelled morphemic segments identified after morphological segmentation. The algorithm presented in this paper corresponds to an intermediary and simple solution since it labels the segments with general morpheme categories, which are detailed in the next section.

4.1 Morpheme Categories

The morphemic segments discovered by the system are labelled by one of the following categories: stem or base (B), prefix (P), suffix (S) and linking element (L). Table 4.1 gives some examples of the labelled morpheme analyses produced by the system compared with the gold standard analyses.

The basic base, prefix and suffix categories are taken into account by several systems which perform unsupervised morphological analysis such as the Morfessor Categories systems [7,8]. The *linking element* category is intended to encompass short segments (usually just one letter long) which link two words or word-forming elements in compounds, such as hyphens, neo-classical linking elements or German *Fugenelemente*. Linking elements differ from the other categories of morphemes because they bring no semantic contribution to the overall meaning of the word.

Table 3. Example morpheme analyses

Word	Method 1	Method 2	Gold standard
Eng. chilly	chill_B y_S	chill_B y_S	chill_A y_s
planners'	planner_B s_S ' _S	plann_B er_S s' _S	plan_N er_s +PL +GEN
Fin. ikuisuus	ikuis_B uus_B	ikuis_B uu_L s_S	ikuinen_A +DA-UUS
resoluutio	resoluutio_B	resoluutio_B	resoluutio_N
Ger. bezwingen	be_P zwing_B en_S	be_P zwing_B e_S n_S	be zwing_V +13PL
risikoloser	risiko_B los_S er_S	risiko_B los_S er_S	risiko_N los +ADJ-er
Tur. avucuna	a_P v_P ucu_B na_S	a_P v_P ucu_B na_S	avuc +POS2S +DAT
kolaCan	kol_B aCan_B	kol_B aCan_B	kolaCan

4.2 Accuracy of Morpheme Labelling

As stated in the introduction, a further objective of morpheme labelling is to disambiguate cases of allomorphy and homography. The recognition of allomorphy is beyond reach of the system in its current state. For instance, the allomorphs of the English prefix *in-* (*im-*, *in-*, *ir-*) or of the suffix *+able* (*-able*, *-ible*) will always be recognised as different morphemes.

Homography should be partially dealt with by the system when homographs belong to different morphemic categories, which excludes within-category homography as *squash_N* and *squash_V* where “squash” will only be identified as a stem.

In order to verify this assumption, let us consider the example of the segment ‘ship’ in English. This segment is either a stem (meaning ‘vessel’) or a suffix which refers to a state. The segment ‘ship’ is correctly labelled as a suffix by method 1 in words like “censorship” (*censor_B ship_S*) or “citizenship” (*citizen_B ship_S*). The stem ‘ship’ can be correctly identified either by method 1 or 2 when it is found at the beginning of a word, but not when it is found at the end of a word; for instance, “shipwreck” is analysed as *ship_B wreck_B* by methods 1 and 2, while “cargo-ship” is analysed as *cargo_B -L ship_S* by method 1.

The previous examples reveal that the simple morpheme labelling performed by the system does not solve all detectable cases of homography between stems and affixes. Morphotactic constraints help in that respect, since they prevent a suffix from occurring at the beginning of a word, and thus the suffix ‘ship’ will not be identified at word initial positions. However, the final analysis privileges the most frequent segment, when several morpheme categories are morphotactically plausible. This tends to be favourable to affixes since they are usually more frequent than stems.

5 Future Work

As shown in the previous section, morphotactic constraints, as they are currently used in the system, are not always sufficient and flexible enough to disambiguate between several homographic segments. For the time being, these constraints are implemented as a simple deterministic automaton, which is the same for all languages. In the future versions of the system, it would be desirable to bootstrap these constraints from the data themselves, as suggested by [9].

Another direction for future research concerns the integration of corpus-derived information. Several algorithms have demonstrated the usefulness of contextual information for unsupervised morphological analysis, to complement orthographic information with semantic and syntactic constraints. Corpus-derived knowledge can be used either at the beginning of the process [10,11], or at the end [12]. In the first case, only words which are contextually similar are compared to discover morphemes. In the second case, spurious morphological analyses are filtered out by taking semantic similarity into account. Corpus-derived information could be incorporated in the first step of the current algorithm, in order to increase the precision of affix acquisition since most of the subsequent processes rely on the affixes

acquired at step 1. Also, it is obviously worth investigating the use of text corpora to achieve finer-grained morpheme labelling and refine the very general categories used so far.

References

1. Kurimo, M., Creutz, M., Varjokallio, M.: Morpho Challenge Evaluation using a Linguistic Gold Standard. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 864–873. Springer, Heidelberg (2008)
2. Kurimo, M., Creutz, M., Turunen, V.: Morpho Challenge Evaluation by IR Experiments. In: Proceedings of the CLEF 2007 Workshop. LNCS. Springer, Heidelberg (2008)
3. Bernhard, D.: Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment. In: Kurimo, M., Creutz, M., Lagus, K. (eds.) Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes, Venice, Italy, pp. 19–23 (April 2006)
4. Déjean, H.: Morphemes as Necessary Concept for Structures Discovery from Un-tagged Corpora. In: Powers, D. (ed.) Proceedings of the CoNLL98 Workshop on Paradigms and Grounding in Language Learning, pp. 295–298 (1998)
5. Hafer, M.A., Weiss, S.F.: Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10, 371–385 (1974)
6. Harris, Z.: From phoneme to morpheme. *Language* 31(2), 190–222 (1955)
7. Creutz, M., Lagus, K.: Induction of a Simple Morphology for Highly-Inflecting Languages. In: Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON), Barcelona, pp. 43–51 (2004)
8. Creutz, M., Lagus, K.: Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In: Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 2005), Espoo, Finland, pp. 106–113 (2005)
9. Demberg, V.: A Language-Independent Unsupervised Model for Morphological Segmentation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 920–927. Association for Computational Linguistics (2007)
10. Bordag, S.: Two-step Approach to Unsupervised Morpheme Segmentation. In: Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes, Venice, Italy, pp. 25–29 (2006)
11. Freitag, D.: Morphology Induction from Term Clusters. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL 2005), Ann Arbor, Michigan, pp. 128–135. Association for Computational Linguistics (2005)
12. Schone, P., Jurafsky, D.: Knowledge-Free Induction of Morphology Using Latent Semantic Analysis. In: Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop, Lisbon, Portugal (September 2000)

Unsupervised and Knowledge-Free Morpheme Segmentation and Analysis

Stefan Bordag

Natural Language Processing Department, University of Leipzig
sbordag@informatik.uni-leipzig.de

Abstract. This paper presents a revised version of an unsupervised and knowledge-free morpheme boundary detection algorithm based on letter successor variety (LSV) and a trie classifier [1]. Additional knowledge about relatedness of the found morphs is obtained from a morphemic analysis based on contextual similarity. For the boundary detection the challenge of increasing recall of found morphs while retaining a high precision is tackled by adding a compound splitter, iterating the LSV analysis and dividing the trie classifier into two distinctly applied classifiers. The result is a significantly improved overall performance and a decreased reliance on corpus size. Further possible improvements and analyses are discussed.

Keywords: Letter successor variety, morpheme boundary detection, morpheme analysis, distributed similarity.

1 Introduction

The algorithm presented in this paper¹ is a revised version of the *letter successor variety* (LSV) based algorithm [2,3,4] described and implemented previously [5,1]. The additional component of morpheme analysis is based on a prototypical implementation described in [6].

The morpheme segmentation algorithm attempts to find morpheme boundaries within word forms. For a given input word form it results in a segmentation into morphs (as opposed to morphemes). It is based on the assumption that any grammatical function is expressed with only a small amount of different affixes. For example, plural is expressed with only five different morphs in German *-en*, *-s*, *-e*, *-er* (and zero).

In essence, the algorithm measures the amount of various letters occurring after a given substring with respect to some context of other words (in this case semantically similar ones), weighting that value according to bi- and tri-gram probabilities and comparing the resulting score to a threshold. Hence it is designed to handle concatenative morphology and it is likely to fail in finding morpheme boundaries in languages with other types of morphology. The algorithm is not rooted in any particular (linguistic) theory of morphology, especially

¹ A recent implementation of this algorithm is available at <http://wortschatz.uni-leipzig.de/~sbordag/>

since such theories tend to omit the fact that morphemes, as their basic units of interest, are not as simply observable as words. The knowledge about where a morph begins and ends is usually assumed to be given a priori.

The present implementation of the morpheme boundary detection consists of three distinct major parts: a compound splitter, a letter successor variety algorithm using contextual similarity of word forms and a trie based machine learning step. Due to the low performance of the LSV based method in splitting longer words, in a pre-processing step a simple compound splitter algorithm is applied. The LSV part is iterated to increase recall with only a moderate loss of precision. The machine learning part (using a trie) is split into two parts, one with high precision and a subsequent one with high recall.

According to an evaluation using the German Celex [7], each change improves the overall performance slightly. Several possibilities of further improvements and analyses are discussed. Any of the major three parts (compound splitter, LSV algorithm, trie classifier) of the described algorithm can be replaced by or merged with a different algorithm, which should facilitate the combination of this algorithm with others.

The morpheme analysis part is based on statistical co-occurrence of the found morphs and subsequent contextual similarity and a basic rule learning algorithm. The rules are then used to find related morphs where groups of related morphs represent a morpheme.

2 Letter Successor Variety

LSV is a measure of the amount of different letters encountered after (or before) a certain substring, given a set of other strings as context. It is possible to use the entire word list as context for each string and its substrings [3,4]. Alternatively, only a specific set of words may be used as context [1], if a method for the selection of relevant words is included. In order to use LSV to find true morpheme boundaries, this set ideally consists of words that share at least one grammatical feature with the input word. For example, if the input word is *hurried*, then relevant words are past tense forms. It is obvious that in such a case the amount of different letters encountered before the substring *-ed* is maximized.

As has been shown earlier [6], using the entire word list for morpheme boundary detection (global LSV) is inferior to using a simulation of semantic similarity (contextual similarity based on comparing statistically significant co-occurrences) of words to find the relevant ones (local LSV). However, the power-law distribution of word frequencies makes it impossible to compute a proper representation of their usage and accordingly compare such words for usage similarity. Hence, local LSV based morpheme boundary detection might have a high precision, but is guaranteed to have a low recall. Another related method, first globally finding the contextually most similar word pairs and then analyzing their differences [8], appears to have even lower recall than the LSV method.

2.1 Trie Classifier

In order to increase the recall of the local LSV method, a machine learning method was proposed. It is based on training a patricia compact trie (PCT) [9] with morpheme segmentations detected by the local LSV methods. The trained trie can then be used to recursively split all words into morphs, irrespective of their frequency.

Training the trie, as depicted in Figure 1, is performed as follows: Each known morpheme boundary is reformulated as a rule: The entire word is the input string, whereas the shorter half of the word (according to the morpheme boundary) is the class to be learned. The trie learns by adding nodes that represent letters of the word to be learned along with increasing the count of the class for each letter (see Figure 1). With multiple boundaries within a single word form, training is applied recursively, taking the outmost and shortest morphs first (from right to left).

Two distinct tries, a **forward-trie** and a **backward-trie** are used to separately learn suffixes and affixes. The decision which trie to use for any given training instance is based on the length of the morphs. The longer half of the word probably contains the stem, whereas the shorter half is used as the class. In the case of the backward-trie, the word itself is reversed.

The classification is applied recursively as well: For an input string both the backward and forward tries are used to obtain the most probable class. This results in up to two identified morpheme boundaries and hence three parts of the original words. Each part is analyzed recursively in the same way as the entire word form until no further classifications can be found.

In the Morpho Challenge 2005 [10], both the local LSV and a subsequent application of the trie learning were submitted separately. As expected, the LSV method had a high precision, but extremely low recall (only 1.9% for Finnish, for example). The application of the trie increased recall, but also lowered precision

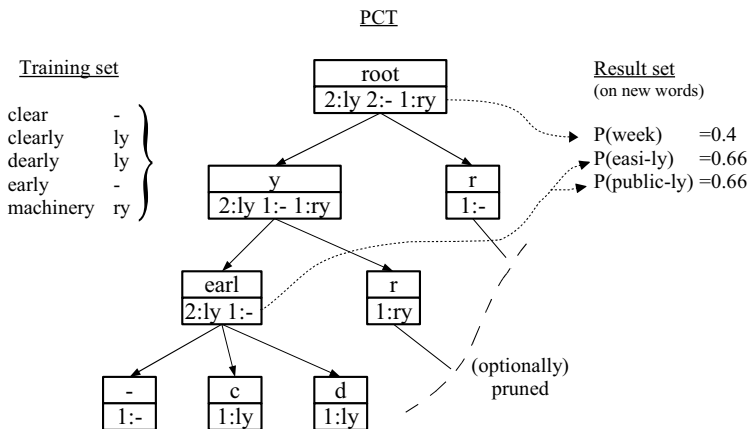


Fig. 1. Illustration of training a PCT and then using it to classify previously unseen words

due to overgeneralization. Overgeneralization occurs mostly because of missing negative examples. Since even for a well-represented input word not all contextually similar words share grammatical information with it, it is impossible to take words without found boundaries as examples of words that in fact do not have any morpheme boundaries.

3 Refined Implementation

The above mentioned weaknesses of the LSV + trie combination hold uniformly over all tested languages. The following modifications attempt to address some of these weaknesses, while trying to avoid language-specific rules or thresholds. The new version contains several changes: a recursive compound identifier, an iteration of the LSV algorithm and splitting the trie classification into two steps.

3.1 Identifying Compounds

The LSV algorithm is based on using contextually similar words. However, compounds usually are less frequent, and words contextually similar to a compound do not necessarily contain other compounds, or compounds sharing parts with the input word. Particularly for semantically opaque compounds this is almost guaranteed to be the case. Therefore it is mostly impossible for the LSV algorithm to find morpheme boundaries between the parts of a compound, unless the compound contains a very productive part.

Since only a small sample set is sufficient for the trie to correctly classify most compounds later, it is not necessary to find all compounds at this point. The compound splitter is therefore based on simply trying to divide a given word *word* at a position *i* and testing whether that division seems plausible. The function *testDiv(word, i)* then tests the plausibility and returns a score. The division is plausible if both parts of the hypothetical decomposition exist as words in the underlying corpus, and reach a threshold of minimum length (4) and a threshold of minimum frequency (20). If that is the case, the score is the sum of the frequencies of the parts assumed to be words.

It is then possible to take the one partition of the input word that maximizes the frequency of the participating parts. This mechanism is applied to recursively divide a long word into shorter units. Table 1 shows that the algorithm (as expected) has a high precision, but it also has a very low recall. In fact, it may have even lower recall for other languages. It also shows that training the trie classifier with this data directly indeed increases recall, but also incurs a rather strong loss in precision. It can be assumed that if compounding exists in a language, then this algorithm in combination with the trie classifier helps to find the parts of a large part of compounds. However, a more elaborate implementation is desirable at this point, especially since this algorithm does not take linking elements into account.

3.2 Iterated LSV Algorithm

For the LSV algorithm, the ideal case is achieved when all contextually similar words to a given input word carry the same grammatical information. However,

due to data sparseness, compounds, overly high co-occurrence frequency and other factors, this ideal state is achieved only for few words. In many cases only a few contextually similar words actually share grammatical information with the input word. Running the LSV algorithm may thus find some morpheme boundaries correctly and not find many others. It is important that in this setup (using contextually similar words as context) it nearly never finds wrong morpheme boundaries, if it does find any, see also Table 1 which shows that the first run of the LSV algorithm found very few (but very precisely) morpheme boundaries.

In order to facilitate the boundary identification for some of the remaining words, it is possible to iterate the LSV algorithm by incorporating knowledge produced in earlier iterations. This is done by an additional factor to the computation of the LSV score: Given a substring *-ly* of the input word *clearly*, if the same substring was identified as a morph in any (or all) of the contextually similar words, then increase the LSV score. However, some very frequent words such as *was* or *do-es* are contextually similar to a large amount of words, which in turn means that these frequent words might influence the analyses of many other words adversely, such as *James* to *Jam-es*. Therefore the increase of the LSV score is normalized against the number of words with the same substring and the number of contextually similar words.

To recall from [6], the formula to compute the left LSV score for the word w at the position i (the formula for the right score is likewise) is:

$$lsv_l(w, i) = plsv_l(w, i) \cdot fw_l(w, i) \cdot ib(w, i) \quad (1)$$

This takes anomalies such as phonemes represented by several letters into account. Here $plsv_l(w, i)$ is the plain number of different letters found to the right of the substring between the beginning of the word w and the position i . $fw_l(w, i)$ is the bi- or trigram based frequency weight of the substring, whereas $ib(w, i)$ is the inverse bigram weight. The previously acquired knowledge about morpheme boundaries is used to compute $prev_l(w, i)$ as the number of previously identified morphs $pf_l(w, i)$ divided by 2 and multiplied with the quotient of the number of words containing the same substring $subf_l(w, i)$ and the size of the pruned list of contextually similar words $prune$:

$$prev_l(w, i) = pf_l(w, i) \cdot 0.5 \cdot (subf_l(w, i)/prune) \quad (2)$$

To prevent the previous analyses from overriding the analysis of the present word, the new LSV score is computed as a multiplication of the LSV score with the previous knowledge, which is at most as high as $lsv_l(w, i) - 1$:

$$lsv2_l(w, i) = \min(lsv_l(w, i) - 1, prev_l(w, i)) \cdot lsv_l(w, i) \quad (3)$$

The same is reversely applied to the right LSV score $lsv_r(w, i)$ and both $lsv_l(w, i)$ and $lsv_r(w, i)$ are summed to produce the final $lsv2(w, i)$ and compare it to a threshold (for example 6) to obtain a decision whether the position i in the word w is a morpheme boundary.

For example, the analyses of the most similar words of *clear-ly* might result in the following morpheme boundaries: *closely*, *white*, *great-ly*, *legal-ly*, *clear*,

linear-ly, really, weakly, ... Hence, for the position 5 (which corresponds to *-ly*) in *clearly*, the amount of previously identified morphs $pf_r(w, i)$ is 3. The number of such substrings $subf_l(w, i)$ is 5 and the amount of contextually similar words was 150. Hence, $prev_r(\textit{clearly}, 5) = 3 \cdot 0.5 \cdot (5/150) = 0.05$ and thus the absolute increase of the LSV score is only 0.05 in this case.

Table 1 shows that there are many cases where the influence was sufficiently strong for the resulting LSV score to reach the threshold. It also shows that iterating the LSV algorithm increases Recall. However, it also incurs a certain Precision loss due with words such as *James* being contextually similar to many other words where *-es* is really a suffix.

Table 1. Iterating the LSV algorithm and applying the modified trie classifier increases recall while keeping precision at high levels

				recursive pretree		
	R	P	F	R	P	F
compounds	10.30	88.33	18.44	27.93	66.45	39.33
lsv_iter_1	17.88	88.55	29.76	57.66	71.00	63.64
lsv_iter_3	23.96	84.34	37.31	62.72	68.96	65.69
saveTrie	31.09	82.69	45.19	66.10	68.92	67.48

3.3 Split Trie Classification

Irrespective of its source, knowledge about boundaries is used to train the trie classifier and then apply the trained classifier to identify more morpheme boundaries. In the original version the trie produces a most probable class for an input string simply by searching for the deepest node in the trie. This mean that often decisions were made without considering further context. For example, the LSV algorithm found the morpheme boundary *drama-tic*. When analyzing *plas-tic*, the trie classifier would find *t* as the deepest matching node. Since that node has only a single class stored with the frequency count of 1, the classifier would decide in favor of *-tic* being a morph with a maximal probability of 1. No further context from the word is considered and the decision is made on grounds of only a single training instance.

However, simply forbidding all decisions that do not take a certain amount of the word into account, would result in extremely low recall, such as 31% for German in Table 1. The trie classification is thus split into two parts, a modified trie classifier and subsequently an original unmodified trie classifier. The modified trie classifier returns a decision only if all of the following conditions are met:

- The deepest matching node must be at least two letters deeper than the class to be returned.
- The matching node must have a minimal distance of three from the root of the trie.
- The total sum of the frequency of all classes stored in the deepest matching node must be larger than 5.

Table 1 shows that applying the modified trie classifier *saveTrie* increases recall by 8% while reducing precision by less than 2%. It also shows that the subsequent application of the original trie classifier further increases recall to a total of 66% while lowering precision to roughly 69%. The table also shows that applying the original trie classifier directly on any of the LSV iterations or even the compound identification algorithm results in lower overall performance.

3.4 Assessing the Improvements

In order to measure the influence of the various improvements proposed, a number of experiments were run on the 3 million sentences German corpus available for the Morpho Challenge 2007. The results of each improvement were measured and are depicted in Table 1. Additionally, the original trie classifier was applied to the results of each modification.

These evaluations show that ultimately, the local LSV implementation could be significantly improved. As such, it reaches similar performance as reported in [1], despite being run on a significantly smaller corpus (3 million sentences vs. 11 million). On the other hand, the relatively small improvements achieved indicate that a significantly better morpheme boundary detection may only be achieved by combining this method with an entirely different approach.

The results of the Morpho Challenge 2007 also show that currently the MDL based approaches to morpheme boundary detection [11,12] mostly outperform the LSV based approach, especially in the more important Information Retrieval task evaluation. The most probable reason is that the LSV algorithm is good at detecting boundaries within high-frequent words, whereas the MDL based algorithms are better at detecting boundaries in longer words. Longer words tend to be less frequent and thus more important for Information Retrieval as opposed to the more frequent words.

A manual analysis of the resulting word list revealed several possible improvements:

- An algorithm specifically designed to identify compounds and take the existence of linking elements into accounts, for example by means of finding reformulations.
- In a post-processing step, an algorithm based on affix signatures such as proposed by [13], might find errors or generalize known morpheme boundaries better than the trie classifiers and ultimately avoid mistakes such as *in-fra-struktur*.
- A global morpheme vocabulary control mechanism, such as the MDL [14,15,11,12] might provide further evidence for or against certain morpheme boundaries and subsequently inhibit mistakes such as *schwa-ech-er*.

4 Morpheme Analysis

Under the assumption that morpheme boundaries were correctly detected, it is possible to treat every single morph separately (similarly to a word) in a

statistical co-occurrence analysis. This allows computing contextual similarity between morphs, instead of words. The following algorithm uses this procedure to find rules that relate various morphs to each other and then applies these rules to produce morphemic analyses of the words that originally occurred in the corpus:

```

for each morph m
  for each cont. similar morph s of m
    if LD_Similar(s,m)
      r = makeRule(s,m)
      store(r->s,m)

for each word w
  for each morph m of w
    if in_store(m)
      sig = createSignature(m)
      write sig
    else
      write m

```

For each morph, the function *LD_Similar(s,m)* filters from the contextually most similar morphs those that differ only minimally, based on Levenshtein Distance (LD) [16] and word lengths. This step could be replaced by a more elaborate clustering mechanism. Pairs with short morphs are only accepted if $LD = 1$, pairs with longer morphs may have a larger distance. The function *makeRule(s,m)* creates a hypothetical rule that explains the difference between two contextually similar morphs. For example, the morphs *ion* and *ions* have a Levenshtein Distance of 1 so the function creates a rule *_-s* (or *n_-ns* to take more context into account) which says that *s* can be added to derive the second morph from the first one. This rule is then stored and associated with the pair of morphs that produced it. This allows deciding between probably correct (if many morph pairs are associated with it) and incorrect rules later.

The second part of the morphemic analysis then applies the acquired knowledge to the original word list. The goal is an analysis of the morphemic structure of all words, where a morpheme is represented by all its allomorphs. In the first step, each word is thus split into its morphs, according to the LSV and trie based algorithm described above. In the next step, all related morphs as stored by the first part of the morphemic analysis are retrieved for each morph of the input word. The function *createSignature(m)* produces a representation of each morpheme. For example, the original word *fracturing* was found to have two morphs: *fractur* and *ing*. The first morph is related to two morphs *fracture* and *fractures*. The second morph is related to *inag*, *ingu* and *iong*. This results in the following analysis:

```

fracturing
> fractur.fracture.fractures
> inag.ing.ingu.iong

```


It is noteworthy that this algorithm cannot distinguish between various meanings of a single morph. In English, the suffix *-s* may be a plural marker if used with a noun or the third person singular marker if used with a verb. Given the extremely high frequency of some of these ambiguous morphs, the number of (at least partially) wrong analyses produced by the algorithm is likely to be high. Further research may evolve around using an unsupervised POS tag inducer [17] to distinguish between different word classes or using a word sense induction algorithm [18] applied to morphs in order to induce the various meanings.

The results from the Morpho Challenge 2007 are surprising in that the morpheme analysis did not yield any significant changes to the evaluation results. This is despite the fact that on average nearly every single morpheme is represented by several morphs. After exploring the word lists for German, the most probable reasons for this appear to be any of the following:

- During construction of the rules no context is taken into account. This often results in morphs to be found as correlated despite them just incidentally looking similar and sharing some contextual similarity. Hence, benefit of the analysis and error might be cancelling each other out.
- Many of the morphs representing a morpheme are, in fact, only artifacts of the mistakes of the morpheme boundary detection algorithm. Thus, the morpheme analysis appears to be strongly influenced by the quality of the detected boundaries.
- When determining the validity of a rule, the amount of morph pairs is taken into account, but not their frequency. This results in many extremely rare morphs (without any impact on the evaluation) to be merged correctly into morphemes, but many very frequent ones (with actual impact on the evaluation) to be missed.

5 Conclusions

Whereas the changes introduced to the morpheme boundary detection improve the overall performance, they also add several more parameters to the entire process. The parameters do not have to be set specifically for each language, but a large number of parameters often indicates the possibility of overfitting. Yet, despite the improvements and the possibility of overfitting, the performance of knowledge-free morpheme boundary detection is far below what knowledge-rich systems (i.e. rule-based) achieve. Nevertheless, the significant beneficial effects achieved in the Information Retrieval evaluation task in the Morpho Challenge 2007 sufficiently demonstrate the usefulness of such algorithms even in the current state.

Compared to other knowledge-free morpheme boundary detection algorithms, the version of the LSV algorithm described in this paper produces good results. The modular design of this algorithm allows for a better interoperability with other algorithms. For example, the significant performance boost achieved by adding a compound splitter indicates that combining various underlying hypotheses is more likely to yield significant improvements than changes to any

single method. Also, given that the most simple combination of algorithms in the form of a voting algorithm in the Morpho Challenge 2005 demonstrated an extraordinary increase in performance, it is reasonable to assume that more direct combinations should perform even better.

The noise produced during the morpheme boundary detection, the missing method for distinguishing ambiguous affixes and other factors resulted in the subsequent morphemic analysis to produce apparently insignificant results. It becomes obvious that adding further algorithmic solutions representing other hypotheses about morpheme boundaries, as well as a more elaborate morphemic analysis, should be a significant step towards a true morphemic analysis similarly to what can be done manually.

References

1. Bordag, S.: Two-step approach to unsupervised morpheme segmentation. In: Proceedings of the PASCAL Challenges Workshop on Unsupervised Segmentation of Words into Morphemes, Venice, Italy (April 2006)
2. Harris, Z.S.: From phonemes to morphemes. *Language* 31(2), 190–222 (1955)
3. Hafer, M.A., Weiss, S.F.: Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10, 371–385 (1974)
4. Déjean, H.: Morphemes as necessary concept for structures discovery from untagged corpora. In: Powers, D. (ed.) Workshop on Paradigms and Grounding in Natural Language Learning at NeMLaP3/CoNLL 1998, Adelaide, Australia, pp. 295–299 (January 1998)
5. Bordag, S.: Unsupervised knowledge-free morpheme boundary detection. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria (September 2005)
6. Bordag, S.: Elements of Knowledge-free and Unsupervised lexical acquisition. PhD thesis, Department of Natural Language Processing, University of Leipzig, Leipzig, Germany (2007)
7. Baayen, R.H., Piepenbrock, R., Gulikers, L.: The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia (1995)
8. Schone, P., Jurafsky, D.: Knowledge-free induction of inflectional morphologies. In: Proceedings of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics, Pittsburgh, PA, USA (2001)
9. Morrison, D.R.: Patricia - practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM* 15(4), 514–534 (1968)
10. Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E., Saraclar, M.: Unsupervised segmentation of words into morphemes - Challenge 2005 An Introduction and Evaluation Report. In: Proceedings of the PASCAL Challenges Workshop on Unsupervised Segmentation of Words into Morphemes, Venice, Italy (2006)
11. Creutz, M., Lagus, K.: Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. In: Publications in Computer and Information Science, Report A81, Helsinki, Finland, Helsinki University of Technology (March 2005)
12. Bernhard, D.: Unsupervised morphological segmentation based on segment predictability and word segments alignment. In: Proceedings of the PASCAL Challenges Workshop on Unsupervised Segmentation of Words into Morphemes (2006)

13. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2), 153–198 (2001)
14. de Marcken, C.: The unsupervised acquisition of a lexicon from continuous speech. Memo 1558, MIT Artificial Intelligence Lab (1995)
15. Kazakov, D.: Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning* 43, 121–162 (2001)
16. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* 163(4), 845–848 (1965)
17. Biemann, C.: Unsupervised part-of-speech tagging employing efficient graph clustering. In: *Proceedings of the Student Research Workshop at the COLING/ACL*, Sydney, Australia (July 2006)
18. Bordag, S.: Word sense induction: Triplet-based clustering and automatic evaluation. In: *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy (April 2006)

Unsupervised Acquiring of Morphological Paradigms from Tokenized Text

Daniel Zeman

Ústav formální a aplikované lingvistiky
Univerzita Karlova
Malostranské náměstí 25
CZ-11800 Praha, Czechia
zeman@ufal.mff.cuni.cz

Abstract. This paper describes a rather simplistic method of unsupervised morphological analysis of words in an unknown language. All what is needed is a raw text corpus in the given language. The algorithm looks at words, identifies repeatedly occurring stems and suffixes, and constructs probable morphological paradigms. The paper also describes how this method has been applied to solve the Morpho Challenge 2007 task, and gives the Morpho Challenge results. Although quite simple, this approach outperformed, to our surprise, several others in most morpheme segmentation subcompetitions. We believe that there is enough room for improvements that can put the results even higher. Errors are discussed in the paper; together with suggested adjustments in future research.

1 Introduction

Morphological analysis (MA) is an important step in natural language processing, needed by subsequent processes such as parsing or translation. Unsupervised approaches to MA are important in that they help process less studied (and corpus-poor) languages, where we have small or no machine-readable dictionaries and tools.

The usual required output of MA is the segmentation of each input word into *morphemes*, i.e. smaller units bearing lexical or grammatical meaning. For instance, the English word *books* would be segmented as *book*+*s*. A supervised morphological analyzer could further put in the information that the meaning of the suffix *s* is “plural”. There is no way how a UMA could learn the label “plural” from an unlabeled text; however, it can learn the segmentation itself, by observing that many English words appear both with and without the *s* suffix.

In many languages, the morphemes are classified as *stems* and *affixes*, the latter being further subclassified as *prefixes* (preceding stems) and *suffixes* (following stems). A frequent word pattern consists of one stem, bearing the lexical meaning, with zero, one or more prefixes (bearing lexical or grammatical meaning) and zero, one or more suffixes (bearing often grammatical meaning). In languages such as German, *compound words* containing more than one stem are quite frequent. While a stem can appear without any affixes, affixes hardly appear on their own, without stems. For the purposes of this paper, a morphological *paradigm* is a collection of affixes that can be attached to the same group of stems, plus the set of affected stems.

Although the segmentation of the word does not provide any linguistically justified explanation of the components of the word, the output can still be useful for further processing of the text. We can recognize stems of new words and possibly replace words by stems. Or we can process morphemes instead of words. Such techniques have been shown to reduce the data sparseness of more complex models like syntactic parsing, machine translation, search and information retrieval.

There is a body of related work that grows faster and faster since the first Morpho Challenge workshop in 2005. [1] first induces a list of 100 most frequent morphemes and then uses those morphemes for word segmentation. His approach is thus not fully unsupervised. [2] extend the ideas of [1]. On the Morpho Challenge 2005 datasets, they achieved the best result for English, but they did remarkably worse for Finnish and Turkish. In contrast, [3] report robust performance with best results for all languages. [4] uses probabilistic distribution of morpheme length and frequency to rank induced morphemes.

The original goal of the present work was to learn paradigms, as defined here. It expects a list of words as input, without actually knowing the frequencies of the words, or knowing how to exploit them. It was tested on Czech, where the most important segment boundary is that between stem and suffix (this is not to say that there are no prefixes or compounds in Czech; there are!) Thus the system assumes that there are only two types of words: atomic (they have only the stem) and two-morpheme words (stem + suffix). This is probably the main weakness of the presented system, and will be addressed later in the discussion.

For the sake of Morpho Challenge, we just ran the paradigm finder over the training corpora, and then searched for the learned stems and suffixes in the test data. There were no attempts (yet) to enrich the system using ideas from the related work. That being said, one might be pleasantly surprised to realize that the system was never the worst one¹ and sometimes even ended above average. This is encouraging, as there clearly are several possible ways of improving the results. We discuss some of them in the concluding section. We leave, however, for the future research to answer whether the system can retain its simplicity while adopting those ideas.

2 Paradigm Acquisition

As said earlier, we do not permit more than one morpheme boundary (i.e. more than two morphemes) in a word.

Example: The word *bank* can be segmented as *bank*, *ban+k*, *ba+nk*, *b+ank*.

There are n possible segmentations of a word of length n , and we iterate over them for each training word. For each stem-suffix pair, we record separately that the suffix was seen with the stem, and that the stem was seen with the suffix. At the end, we have for each suffix a list of stems with which they were seen. We group together suffixes with exactly the same sets of stems. The set of suffixes in the group, plus the set of stems they share, is an (unfiltered) paradigm.

¹ This does not apply to the information retrieval task, where our system occupied the worst rank in most rankings.

2.1 Filtering

The list of paradigms obtained so far is huge and redundant. For instance, if all suffixes in a paradigm begin with the same letter, there is another paradigm which differs only in that the letter has been shifted to the stem. The following example is from Finnish:

Paradigm A

Suffixes: a, in, ksi, lla, lle, n, na, ssa, sta

Stems: erikokoisi funktionaalisi logistisi mustavalkoisi objektiivisi ...

Paradigm B

Suffixes: ia, iin, iksi, illa, ille, in, ina, issa, ista

Stems: erikokois funktionaalis logistis mustavalkois objektiivis ...

Paradigm C

Suffixes: sia, siin, siksi, silla, sille, sin, sina, sissa, sista

Stems: erikokoi funktionaali logisti mustavalkoi objektiivii ...

Paradigm D

Suffixes: isia, isiin, isiksi, isilla, isille, isin, isina, isissa, isista

Stems: erikoko funktionaal logist mustavalko objektiiv ...

We have to filter the paradigms in order to make them useful. We apply the following filtering rules:

2.1.1 More Suffixes Than Stems

Both stem and suffix can be as short as one character. Then how do we recognize that a paradigm with one stem *s* and tens of thousands of suffixes is untrustworthy? We consider suspicious all paradigms where there are more suffixes than stems. Those paradigms are discarded without compensation.

2.1.2 Uniform Letter on the Stem-Suffix Border

As in the Finnish example above, with a uniform letter (or group of letters) on the stem-suffix boundary, we get a set of matching paradigms where the letter(s) is on one or the other side of the boundary. Unlike in the Finnish example, we are not always guaranteed that the corresponding Paradigm B actually does not contain other stems or suffixes, which make the projection irreversible. Example (from Czech):

Paradigm A

Suffixes: l, la, li, lo, ly

Stems: kouři nosi pádi

Paradigm B

Suffixes: il, ila, ili, ilo, ily, ů

Stems: kouř nos pás

In this case, the second paradigm adds the suffix *ů* to the bag, which means that we could not induce Paradigm A from B. On the other hand, the Paradigm B cannot contain additional stems. Consider, for instance, adding a new stem *udobř* to Paradigm B (and removing the *ů* suffix). It would mean that there is a word *udobřil* in the training data. One of the possible segmentations of that word is *udobři-l*, and the same can be done with all the other suffixes, thus we must have had the stem *udobři* in Paradigm A. But we did not.

Similarly, we can proceed from the longer suffixes to the shorter ones. When all suffixes begin with the same letter, there must be a corresponding Paradigm B, where the letter is shifted to the stems. The Paradigm B can contain additional stems, as in the following example:

Paradigm A

Suffixes: il, ila, ili, ilo, ily

Stems: kouř nos pád

Paradigm B

Suffixes: l, la, li, lo, ly

Stems: kouři nosi pádi sedě

While Paradigm B can add stems, it cannot add suffixes. Consider adding a suffix *t* (and removing the stem *sedě*). It would mean that the words *kouřit*, *nosit*, *pádit* were in the training data, and thus the suffix *it* should have appeared in Paradigm A.

Now it is obvious that the boundary letters create room for paradigm filtering. The question is, should we prefer longer stems, or longer suffixes? We decided to prefer longer stems. If all suffixes in a paradigm begin with the same letter, we discard the paradigm, being sure that there is another paradigm with those border letters in stems. That other paradigm may contain some other stems as well, which further strengthens our conviction that the border letter was not a genuine part of the suffixes.

2.1.3 Subsets of Paradigms

A frequent problem is that stems have not been seen with all applicable suffixes. Consider the following example (from real Czech data):

A.suffixes = {ou, á, é, ého, ém, ému, ý, ých, ým, ými}

B.suffixes = {ou, á, é, ého, ém, ému, ý, ých, ým}

C.suffixes = {ou, á, é, ého, ém, ý, ých, ým, ými}

D.suffixes = {ou, á, é, ého, ém, ý, ých, ým}

As a matter of fact, stems of all four paradigms should belong to the paradigm A but not all of them occurred with all A suffixes. As one important motivation of UMA is to cover unknown words, it is desirable to merge the subset paradigms with their superset A. Unfortunately, this can sometimes introduce stem+suffix combinations that are not permitted in the given language.

When talking about set inclusion on paradigms, we always mean the sets of suffixes, not stems. If the suffixes of Paradigm B form a subset of suffixes of Paradigm A, and there is no C, different from A, such that B is also subset of C, then merge A with B (which means: keep suffixes from A, and stems from both).²

The implementation of this rule is computationally quite complex. In order to identify subset relations, we would have to step through n^2 paradigm pairs (n is the current number of paradigms, over 60,000 for our Czech data), and perform k comparisons for each pair (in half of the cases, k is over 5). As a result, tens of billions of comparisons would be needed.

That is why we do not construct the complete graph of subsets. We sort the paradigms with respect to their size, the largest paradigm having size (number of suffixes)

² If the other superset C exists, it is still possible that the merging will be enabled later, once we succeed to merge A with C.

k. We go through all paradigms of the size $k-1$ and try to merge them with larger paradigms. Then we repeat the same with paradigms of the size $k-2$, and so on till the size 1. The total number of comparisons is now much lower, as the number of paradigms concurrently decreases.

For each paradigm, we check only the closest supersets. For instance, if there is no superset larger by 1, and there are two supersets larger by 2, we ignore the possibility that there is a superset larger by 3 or more. They are linked from the supersets larger by 2. If an ambiguity blocks simplifying the tree, it is not a reason to block simplifying on the lower levels.

2.1.4 Single Suffix

Paradigms with a single suffix are not interesting. They merely state that a group of words end in the same letters. Although we could identify unknown words belonging to the same group and possibly segment them along the border between the non-matching and matching part, there is not much to be gained from it. There is also no guarantee that the matching end of the word is really a suffix (consider a paradigm with suffix *n* and “stems” from thousands of words ending in *n*). So we discard all single-suffix paradigms and thus further simplify the paradigm pool.

2.2 More Paradigm Examples

For illustration, we provide some of the largest (with most suffixes) paradigms for the four languages of Morpho Challenge 2007 and Czech:

English

e, ed, es, ing, ion, ions, or
calibrat consecrat decimat delineat desecrat equivocat postulat regurgitat
0³, d, r, r's, rs, s
analyze chain-smoke collide customize energize enquire naturalize scuffle ...

Finnish

0, a, an, ksi, lla, lle, n, na, ssa, sta, t
asennettava avattava hinattava koordinoiva korvattava leijuva mahdollistama ...
a, en, in, ksi, lla, lle, lta, na, ssa, sta
ammattinharjoittaji avustavi jakavi muuttaji omaavi parannettavi puolueettomi ...

German

0, m, n, r, re, rem, ren, rer, res, s
aggressive bescheidene deutliche dunkle flexible langsame mächtige ruhige ...
0, e, em, en, er, es, keit, ste, sten
entsetzlich gutwillig lebensfeindlich massgeblich reichhaltig unbarmherzig ...

Turkish

0, de, den, e, i, in, iz, ize, izi, izin
anketin becerilerin birikimlerin gereksinimin giysilerin görüntülerin güvenin ...
0, dir, n, nde, ndeki, nden, ne, ni, nin, yle
aleti arabirimi etiketi evreleri geçilmesi geçişleri iletimi iliği kanseri ...

³ 0 means empty suffix.

Czech

ou, á, é, ého, ém, ému, ý, ých, ým, ými
gruzínsk italsk lékařsk ministersk městsk někter olympijsk poválečn pražsk ...
0, a, em, ovi, y, ů, ům
divák dlužník obchodník odborník poplatník právník předák vlastník útočník ...

3 Segmenting a Word

Given a set of paradigms for a language, how do we apply it to segment a word in that language? Actually, we only use the sets of all stems and all suffixes in the Morpho Challenge task. We do not exploit the information that a stem and a suffix occurred in the same paradigm. Yet the acquisition of paradigms described in the previous section is still important, as it greatly reduces the number of learned stems and suffixes.

Again, we consider all possible segmentations of each analyzed word. For each stem-suffix pair, we look up the table of learned stems and suffixes. If both stem and suffix are found, we return that particular segmentation as a possible analysis. (Note that more than one segmentation can satisfy the condition, and thus ambiguous analyses are possible.) If no analysis is found this way, we return analyses with known suffixes or known stems (but not both). If no analysis is found either way, we return the atomic analysis, i.e. the entire word is a stem, the suffix is empty.

4 Results

The Morpho Challenge 2007 task does not (and actually cannot) require that the morphemes in segmentation be labeled in any particular way. Due to possible phonological changes caused by inflection of words, the segmenters are not even required to denote the exact position of the morpheme border in the word. Therefore, the only information that can be compared with a gold standard is the number of morphemes in the word, and the fact that two words share a morpheme with the same label on specified positions. The precise description of the evaluation algorithm is available at the Morpho Challenge website.⁴ We present only the results of the Competition 1 in this paper.⁵

In comparison to other systems, our system usually did better w.r.t. recall than w.r.t. precision. The best rank achieved by our system was for Turkish, while for the other languages we ended up below average. For each language, we provide our rank and the number of ranked systems (in addition to the percentages). P is precision, R is recall, F is their harmonic mean.

The processing of each language took from several minutes to several hours. Finnish needed the most time due to its enormous agglutinative morphological system. German is not an agglutinative language but its long compound words also increased

⁴ <http://www.cis.hut.fi/morphochallenge2007/evaluation.shtml>

⁵ In the Competition 2, the segmentation results are evaluated indirectly by using them in an information retrieval task. Our system was among the poorest in all rankings of the Competition 2 but we have currently no plausible explanation.

	English		
	P	R	F
%	52.98	42.07	46.90
rank	10	5	9
# ranked	13		

	German		
	P	R	F
%	52.79	28.46	36.98
rank	9	9	9
# ranked	12		

	Finnish		
	P	R	F
%	58.84	20.92	30.87
rank	8	6	6
# ranked	9		

	Turkish		
	P	R	F
%	65.81	18.79	29.23
rank	8	2	2
# ranked	9		

time requirements. Turkish was faster not because it is less complex but simply because of the much smaller data set.

Not surprisingly, the slowest part of the algorithm is the subset pruning.

5 Discussion

The presented approach is a truly unsupervised one, as it does not need any language-specific tuning ever (compare with the lists of most frequent morphemes in some related work). However, there are many lessons to be learned from other systems and tested during future research. Some ideas follow:

- Our system does not (but it should) exploit the word/morpheme frequencies in the corpus. Very rare words could be typos and could introduce nonsensical morphemes.
- Our system reduces morpheme segmentation to just one stem (mandatory) and one suffix (optional). Such limitation is too severe. At least we ought to enable prefixes. It could be done by repeating the process described in this paper, but now the second part would be a stem and the first part an affix. Using the new model, we could recognize prefixes in the stems of the old model, and using the old model, we could recognize suffixes in the new one.
- Even that is fairly limited. There are composite suffixes (as in English *compose+r+s*) and composite prefixes (as in German *ver+ab+schieden*). A good morphological analyzer should identify them (not to mention that they are likely to appear in the gold standard data).
- Finally, compounds make the possible number of morphemes virtually unlimited. (An anecdotic German example is *Hotentot + en + potentat + en + tante + n + atentät + er*.) A possible partial solution is to do a second run through the stems and identify combinations of two or more smaller stems. However, as seen from the example, suffixes are involved in compound creation as well.
- Morphological grammars of many languages contain rules for phonological changes (for instance, *deny* vs. *deni* in English *denial*, Czech *matk+a*, *matc+e*, *matč+in*, German *Atentat* vs. *Atentät+er*). Supervised MA systems have

incorporated such rules in order to succeed (e.g., see [5] or [6]). [3] induce phonological rules for suffixes longer than 1 character, however, the above Czech example suggests that it may be needed for suffixes of length 1 as well.

6 Conclusion

We have presented a paradigm acquisition method that can be used for unsupervised segmentation of words into morphemes. The approach is very simple; however, even such a simple system turned out to be reasonably successful. It gives us the hope that by incorporating the ideas from Discussion, we can catch up with at least some of the better systems from Morpho Challenge 2007.

Acknowledgements. This research has been supported by the Czech Academy of Sciences, the “Information Society” project No. 1ET101470416, and Ministry of Education of the Czech Republic, project MSM0021620838.

References

1. Déjean, H.: Morphemes as Necessary Concepts for Structures Discovery from Untagged Corpora. In: Worksh. on Paradigms and Grounding in Nat. Lang. Learning, pp. 295–299 (1998)
2. Keshava, S., Pitler, E.: A Simple, Intuitive Approach to Morpheme Induction. In: PASCAL Challenge Works. on Unsup. Segm. of Words into Morphemes, Southampton (2006)
3. Dasgupta, S., Ng, V.: High-Performance, Language-Independent Morphological Segmentation. In: Proc. of NAACL HLT, Rochester, pp. 155–163 (2007)
4. Creutz, M.: Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. In: Proc. of ACL, Sapporo (2003)
5. Koskeniemi, K.: Two-level Morphology: A General Computational Model for Word-form Recognition and Production. Pub. No. 11. U. of Helsinki, Dept. of Gen. Ling (1983)
6. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Univerzita Karlova, MFF, Ústav formální a aplikované lingvistiky. Praha (2004)

ParaMor: Finding Paradigms across Morphology^{*}

Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin

Language Technologies Institute
Carnegie Mellon University
{cmonson, jgc+, alavie+, lsl+}@cs.cmu.ed

Abstract. ParaMor automatically learns morphological paradigms from unlabelled text, and uses them to annotate word forms with morpheme boundaries. ParaMor competed in the English and German tracks of Morpho Challenge 2007 (Kurimo et al., 2008). In English, ParaMor’s balanced precision and recall outperform at F_1 an already sophisticated baseline induction algorithm, Morfessor (Creutz, 2006). In German, ParaMor suffers from a low morpheme recall. But combining ParaMor’s analyses with analyses from Morfessor results in a set of analyses that outperform either algorithm alone, and that place first in F_1 among all algorithms submitted to Morpho Challenge 2007.

Categories and Subject Descriptions: I.2 [Artificial Intelligence]: I.2.7 Natural Language Processing.

Keywords: Unsupervised Natural Language Morphology Induction, Paradigms.

1 Introduction

This paper describes ParaMor, an algorithm that automates the construction of a morphology analysis system for any language from nothing more than unannotated text. We present and discuss ParaMor’s performance in Morpho Challenge 2007 (Kurimo et al., 2008), a competition for unsupervised algorithms that induce the morphology of natural languages.

Following both traditional and modern theories of inflectional morphology (Stump, 2001), our work on unsupervised morphology induction recognizes the paradigm as the natural organizational structure of inflectional morphology. A paradigm is a set of surface forms that a lexeme can take as it inflects for relevant morphosyntactic features. For example *bueno*, *buenos*, *buen*, *buenas* is the Spanish paradigm for gender and number on adjectives. ParaMor exploits paradigms by identifying sets of mutually exclusive strings which closely align with natural language paradigms, and then segmenting word forms into morpheme-like pieces suggested by the discovered paradigms. Currently, ParaMor can isolate word final suffixes. ParaMor’s methods can be straightforwardly generalized to prefixes and forthcoming work models sequences of concatenative morphemes.

^{*} The research reported in this paper was funded in part by NSF grant number IIS-0121631.

Previously proposed minimally supervised approaches to the induction of morphology have also drawn on the unique structure of natural language morphology. Emphasizing morphemes as recurrent building blocks of words, Brent et al. (1995), Goldsmith (2001), and Creutz (2006) each use recurring word segments to efficiently encode a corpus. These approaches then hypothesize that those recurring segments which most efficiently encode a corpus are likely morphemes. Another technique that exploits morphemes as repeating sub-word segments encodes the lexemes of a corpus as a character tree, i.e. a trie, (Harris, 1955; Hafer and Weis, 1974; Demberg, 2007), or as a finite state automaton (FSA) over characters (Johnson, H. and Martin, 2003; Altun and M. Johnson, 2001). A trie or FSA conflates multiple instances of a morpheme into a single sequence of states. The paradigm structure of natural language morphology has also been previously leveraged. Goldsmith (2001) uses morphemes to efficiently encode a corpus, but he first groups morphemes into paradigm like structures he calls signatures, while Snover (2002) incorporates paradigm structure into a generative statistical model of morphology.

2 ParaMor

We present our unsupervised morphology induction algorithm, ParaMor, by following an extended example of the analysis of the Spanish word *administradas*, the *feminine plural past participle* form of *administrar* ‘to administer’. The word *administradas* occurs in the corpus of Spanish newswire on which we developed the ParaMor algorithm. This Spanish newswire corpus contains 50,000 types. We hope the detailed example we give here can flesh out the abstract step-by-step description of ParaMor in Monson et al. (2007).

Before delving into ParaMor’s details we note two facts which guided algorithm design. First, in any given corpus, a particular lexeme will likely not occur in all possible inflected forms. But rather each lexeme will occur in some subset of its possible surface forms. Second, we expect inflected forms of a single lexeme to be correlated. That is, if we have observed several lexemes in inflected form *A*, and if *B* belongs to the same paradigm as *A*, then we can expect a significant fraction of those lexemes inflected as *A* to also occur in inflected form *B*.

Search: ParaMor begins with a search for partial paradigms, where a partial paradigm is a set of candidate suffixes, and a candidate suffix is any final substring of any word in the corpus. The word *administradas* gives rise to many candidate suffixes including: *stradas*, *tradas*, *radas*, *adas*, *das*, *as*, *s*, and $\emptyset$. The candidate suffix *s* is a true morpheme of Spanish, marking plural. Additionally, the left edges of the word-final strings *as* and *adas* occur at Spanish morpheme boundaries. Of course, while we can discuss which candidate suffixes are reasonable and which are not, ParaMor, as an unsupervised morphology induction system, has no a priori knowledge of Spanish morphology.

Any particular candidate suffix may be derived from multiple word forms. The (incorrect) candidate suffix *stradas* occurs as the final substring of eight wordforms in our Spanish corpus, including the words *administradas*, *arrastradas* ‘wretched’ and *mostradas* ‘accustomed’. The candidate suffix *s* is a word final string of 10,662 wordforms in this same corpus—more than one fifth of the unique wordforms! When a

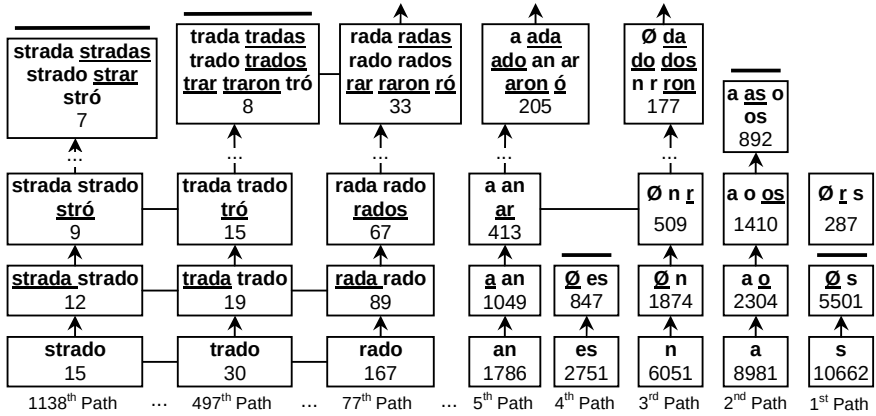


Fig. 1. Eight paths ParaMor follows in search of likely partial paradigms. Search paths begin at the bottom of the figure and follow the arrows upward. A horizontal bar indicates the termination of a search path. Candidate suffixes appear in **bold**. Underlined candidate suffixes have been added by the most recent search step(s). Each partial paradigm gives the number of candidate stems which attach to all candidate suffixes in that partial paradigm. Horizontal links between partial paradigms connect sets of suffixes that differ only in their initial character.

candidate suffix is stripped from a surface word, we call the remaining word initial string a candidate stem. The (incorrect) candidate suffix *stradas* gives rise to eight (incorrect) candidate stems including *admini*, *arra*, and *mo*.

ParaMor’s initial search for partial paradigms considers every candidate suffix derived from any word form in the input corpus as potentially part of a true inflectional paradigm. ParaMor’s search considers each non-null candidate suffix in turn, beginning with that candidate suffix which can attach to the most candidate stems, working toward suffixes which can attach to fewer stems. For each particular candidate suffix, f , ParaMor notes the candidate stems, T , to which f can attach, and then identifies the candidate suffix, f' , that forms separate corpus words with the largest number of stems in T . The candidate suffix f' is then added to the partial paradigm anchored by f . Now with a partial paradigm containing two candidate suffixes, ParaMor resets T to be the set of candidate stems which form corpus words with both f and f' . ParaMor then searches for a third suffix which can form words with the largest subset of this new T . ParaMor continues to add candidate suffixes until one of two halting criteria is met:

1. Since we expect suffixes from a single paradigm to be correlated, ParaMor stops growing a partial paradigm if no candidate suffix can form corpus words with at least a threshold fraction of the stems in the current partial paradigm.
2. ParaMor stops adding candidate suffixes if the stem evidence for the partial paradigm is too meager—ParaMor will only add a suffix to a partial paradigm if there are more stems than there are suffixes in the proposed partial paradigm.

Fig. 1 contains portions of a number of search paths that ParaMor followed when analyzing our Spanish newswire corpus. Most of the paths in Fig. 1 are directly relevant to the analysis of *administradas*. Search paths begin at the bottom of Fig. 1 and

proceed upwards. In Spanish, the non-null candidate suffix that can attach to the most stems is *s*. The search path begun from *s* is the right-most search path shown in Fig. 1. The null suffix, $\emptyset$, can attach to the largest number of candidate stems to which *s* can attach, and so the first search step adds $\emptyset$ to the candidate suffix *s*. ParaMor then identifies the candidate suffix *r* as the suffix which can attach to the most stems to which *s* and $\emptyset$ can both attach. But *r* can only form corpus words in combination with 287 or 5.2% of the 5501 stems to which *s* and $\emptyset$ can attach. As such a severe drop in stem count does not convincingly suggest that the candidate suffix *r* is correlated with the candidates *s* and $\emptyset$, ParaMor does not add *r*, or any other suffix, to the now closed partial paradigm *s.∅*. Experimentally we determined that, for Spanish, requiring at least 25% of stems to carry over when adding a candidate suffix discovers reasonable partial paradigms. The left-most search path in Fig. 1, which begins from *strado*, illustrates the second stopping criterion. From *strado* four candidate suffixes are added one at a time: *strada*, *stró*, *strar*, and *stradas*. Only seven candidate stems form words when combined singly with all five of these candidate suffixes. Adding any additional candidate suffix to these five suffixes brings the stem count down at least to six. Since six stems is not more than the six suffixes which would be in the resulting partial paradigm, ParaMor does not add a sixth candidate suffix.

In our corpus of Spanish newswire text, ParaMor's initial search identifies partial paradigms containing 92% of all inflectional suffixes of Spanish, or 98% of inflectional suffixes that occurred at least twice in the corpus. Among the selected partial paradigms are those which contain portions of all nine inflectional paradigms of Spanish. The high recall of the initial search comes, of course, at the expense of precision. While our analysis provides nine true paradigms containing 87 unique suffixes, ParaMor constructs 8339 partial paradigms with 9889 unique candidate suffixes.

The constructed partial paradigms have three flaws. First, the candidate suffixes of many partial paradigms overlap. At the end of the initial search, there are 27 distinct partial paradigms that contain the reasonable candidate suffix *adas*. Each of these 27 partial paradigms comes from a distinct initial candidate suffix: *an*, *en*, *ación*, *amos*, etc. The second flaw is that most constructed partial paradigms contain many fewer candidate suffixes than do the true paradigms of Spanish. And third, many partial paradigms include candidate suffixes with incorrect morpheme boundaries. ParaMor addresses the first two flaws by merging together similar partial paradigms. And ParaMor addresses the third flaw, while further ameliorating the second, through filters which weed out less likely paradigm clusters.

Clustering: To merge partial paradigms, ParaMor adapts greedy hierarchical agglomerative clustering. Fig. 2 contains a small portion of the partial paradigm cluster that consumes the partial paradigm built from the candidate suffix *an*. Part of the search path from *an* is summarized in Fig. 1. But the search path continues until there are fifteen candidate suffixes in the partial paradigm: *a*, *aba*, *aban*, *ada*, *adas*, *ado*, *ados*, *an*, *ando*, *ar*, *aron*, *arse*, *ará*, *arán*, and *ó*. The partial paradigm built from *an* appears on the center right of Fig. 2. During clustering, *an*'s partial paradigm is merged with a cluster that has previously formed. This previously formed cluster and the two partial paradigms which merged to form it appear at the bottom left of Fig. 2. ParaMor decides which partial paradigm clusters to merge by computing a similarity score between pairs of paradigm clusters. A variety of similarity metrics on partial paradigms are possible. Looking at Fig. 2, it is clear that both the candidate suffix sets



Fig. 2. A portion of a cluster of partial paradigms. The candidate suffixes of each partial paradigm or cluster node appear in **bold**, candidate stems are in *italics*. Suffixes in cluster nodes which uniquely originate in one child are underlined.

and the candidate stem sets of partial paradigms can overlap. Consequently partial paradigms can share covered surface types. For example, the bottom two clusters of Fig. 2 both contain the candidate suffix *a* and the candidate stem *anunci*. Reconcatenating this stem and suffix we say that both of these partial paradigms cover the boundary annotated word form *anunci+a*. ParaMor computes the similarity of partial paradigms, and their clusters, by comparing just such sets of morpheme boundary annotated word forms. We have found that the particular similarity metric used does not significantly affect clustering. For the experiments we report here we use the cosine similarity for sets, given as $|X \cap Y| / (\|X\| \|Y\|)^{1/2}$. It is interesting to note that similarity scores do not monotonically decrease moving up the tree structure of a particular cluster. Non-decreasing similarities is a consequence of computing similarities over sets of objects which are merged up the tree. Returning to our Spanish example word *administradas*, clustering reduces, from 27 to 6, the number of distinct partial paradigms in which the candidate suffix *adas* occurs. Clustering also reduces the total number of separate partial paradigms to 7511 from 8339.

Filtering: With the fragmentation of partial paradigms significantly reduced, ParaMor focuses on removing erroneously proposed partial paradigm clusters. The first filtration step removes all partial paradigms which do not cover at least a threshold number of word forms. Monson et al. (2007) discusses our empirical procedure to identify a reasonable threshold. This first filter drastically reduces the number of selected partial paradigms, from 7511 to 137. Among the many discarded partial paradigms is one of the six remaining partial paradigms containing *adas*. Although *adas* can be a valid verbal suffix sequence, the discarded partial paradigm was built from forms including *gradas* ‘stairs’ and *hadas* ‘fairies’, both nouns. Also removed are all partial paradigms containing the incorrect candidate suffix *stradas*.

Of the 137 remaining partial paradigm clusters, more than a third clearly attempt to model a morpheme boundary to the left of a correct morpheme boundary. Among

these left-leaning clusters are those containing the partial paradigms built from the candidate suffixes *trado* and *rado*, given in Fig. 1. To filter out left-leaning clusters, ParaMor adapts a strategy due to Harris (1955), who detects morpheme boundaries by examining word internal character variation. Consider the partial paradigm *trada-tradas.trado.trados.trar.traron.tró*, in which all seven candidate suffixes begin with *tr*. In Fig. 1 this *tr*-paradigm is linked to the right with the partial paradigm *rada-radas.rado.rados.rar.raron.ró*, obtained by removing the initial *t* from each candidate suffix. Although not pictured in Fig. 1, the *r*-paradigm is further connected to the partial paradigm *ada.adas.ado.ados.ar.aron.ó* through removal of the initial *r*. Because the stems of this *ada*-containing partial paradigm exhibit a wide variety of final characters, 19 in all, ParaMor hypothesizes that the correct morpheme boundary is *after* the *tr* sequence—And ParaMor removes the *trada*-containing partial paradigm. We measure the stem final character variety within a partial paradigm using entropy. If stem final character entropy falls above a threshold value then ParaMor takes that partial paradigm as modeling a morpheme boundary. We have found that even a conservative, low, entropy cutoff discards nearly all clusters which model a morpheme boundary too far to the left. Applying this filter leaves 80 clusters, and furthermore completely removes any and all clusters containing either candidate suffixes *tradas* or *radas*. ParaMor currently contains no method for discarding clusters which place a morpheme boundary to the right of the correct position.

Segmentation: Finally, with a strong grasp on the paradigm structure, ParaMor segments the words of a corpus into morphemes by stripping off suffixes which likely participate in a paradigm. ParaMor’s segmentation algorithm is easily described by finishing out our extended example of the analysis of the word *administradas*. Among the 80 paradigm clusters that ParaMor accepts are clusters containing the candidate suffixes *adas*, *das*, *as*, and *s*. Of these, *adas*, *as*, and *s* identify correct morpheme boundaries. The clusters containing the candidate suffix *das* cannot be removed with either the size filter or the currently implemented morpheme boundary filter. Among the clusters which contain *adas* several also contain *ada*; similarly *das* and *da*, *as* and *a*, and *s* and $\emptyset$, each appear together in at least one cluster. Replacing, in *administradas*, *adas* with *ada*, *das* with *da*, *as* with *a*, or *s* with $\emptyset$ results in the potential word form *administrada*—a form which occurs in our Spanish corpus. Using this information, ParaMor produces four separate analyses of *administradas*: *administr +adas*, *administra +das*, *administrad +as*, and *administrada +s*.

3 Morpho Challenge 2007 Results and Conclusions

We entered ParaMor in the English and the German tracks of Morpho Challenge 2007. In each track we submitted three systems. The first system was ParaMor alone. We did not vary ParaMor’s free parameters, but held each at a setting which produced reasonable *Spanish* suffix sets (Monson et al., 2007). The English and German corpora used in Morpho Challenge 2007 were larger than we had previously worked with. The English corpus contains nearly 385,000 types, while the German corpus contains more than 1.26 million types. ParaMor induced paradigmatic scheme-clusters over these larger corpora from just the top 50,000 most frequent types. But

with the scheme-clusters in hand, ParaMor segmented all the types in each corpus. The second submitted system combines the analyses of ParaMor with the analyses of Morfessor (Creutz, 2006). We downloaded Morfessor Categories-MAP 0.9.2 (Creutz, 2007) and optimized Morfessor’s single parameter separately for English and for German. We optimized Morfessor’s parameter against an F_1 score calculated following the methodology of Morpho Challenge 2007. The Morpho Challenge F_1 score is found by comparing Morfessor’s morphological analyses to analyses in human-built answer keys. The official Morpho Challenge 2007 answer keys were not made available to the challenge participants. However, the official keys for English and German were created using the Celex database (Burnage, 1990), and Celex was available to us. Using Celex we created our own morphological answer keys for English and German that, while likely not identical to the official gold standards, are quite similar. Optimizing Morfessor’s parameter renders the analyses we obtained from Morfessor no longer fully unsupervised. In the submitted combined system, we pooled Morfessor’s analyses with ParaMor’s straightforwardly: for each analyzed word, we added Morfessor’s analysis as an additional, comma separated, analysis to the list of analyses ParaMor identified. Naively combining the analyses of two systems in this way increases the total number of morphemes in each word’s analyses—likely lowering precision but possibly increasing recall. The third set of analyses we submitted to Morpho Challenge 2007 is the set Morfessor produced alone at the same optimized parameter settings used in our combined entry.

Table 1 contains the official Morpho Challenge 2007 results for top placing systems in English and German. In English, ParaMor’s more balanced precision and recall outperform, at F_1 , the baseline Morfessor system with its precision centric analyses. As expected, combining ParaMor’s and Morfessor’s analyses boosts recall but hurts precision. The net effect on F_1 in English is a negligible improvement over ParaMor alone. In German, however, ParaMor’s precision is significantly higher than for English. And combining analyses retains a respectable overall precision. The unified ParaMor-Morfessor system achieved the highest F_1 of any submitted system. Bernhard is a close second just 0.3 absolute lower—a likely statistically insignificant difference.

Table 1. The official Precision, Recall, and F_1 scores from Morpho Challenge 2007, to three significant digits. Only scores for submitted systems most relevant to a discussion of ParaMor are included. * The Submitted System Morfessor is that trained by Monson et al.

Submitted Systems	English			German		
	P	R	F_1	P	R	F_1
ParaMor & Morfessor	41.6	65.1	50.7	51.5	55.6	53.2
ParaMor	48.5	53.0	50.6	59.1	32.8	42.2
Morfessor*	77.2	34.0	47.2	67.2	36.8	47.6
Bernhard-2	61.6	60.0	60.8	49.1	57.4	52.9
Bernhard-1	72.1	52.5	60.7	63.2	37.7	47.2
Pitler	74.7	40.6	52.3	N/A	N/A	N/A
Bordag-5a	59.7	32.1	41.8	60.5	41.6	49.3
Zeman	53.0	42.1	46.9	52.8	28.5	37.0

We are excited by ParaMor's strong performance and eager to extend our algorithm. Recent experiments suggest the precision of ParaMor's segmentations can be improved by building partial paradigms from cleaner data. Perhaps ParaMor and Morfessor's complementary morphological analyses can be combined in an even more fruitful fashion. And ongoing work addresses affix sequences by merging ParaMor's multiple distinct analyses.

References

- Altun, Y., Johnson, M.: Inducing SFA with e-Transitions Using Minimum Description Length. In: *Finite State Methods in Natural Language Processing Workshop at ESSLLI*, Helsinki, Finland (2001)
- Brent, M.R., Murthy, S.K., Lundberg, A.: Discovering Morphemic Suffixes: A Case Study in MDL Induction. In: *The Fifth International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, Florida (1995)
- Burnage, G.: *Celex—A Guide for Users*. Springer, Centre for Lexical information, Nijmegen, The Netherlands (1990)
- Creutz, M.: Morpho project (May 31, 2007), <http://www.cis.hut.fi/projects/morpho/>
- Creutz, M.: Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition. Ph.D. Thesis. Computer and Information Science, Report D13. Helsinki: University of Technology, Espoo, Finland (2006)
- Demberg, V.: A Language-Independent Unsupervised Model for Morphological Segmentation. Association for Computational Linguistics, Prague (2007)
- Goldsmith, J.: Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2), 153–198 (2001)
- Hafer, M.A., Weiss, S.F.: Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval* 10(11/12), 371–385 (1974)
- Harris, Z.: From Phoneme to Morpheme. *Language* 31(2), 190–222 (1955); Reprinted in Harris (1970)
- Harris, Z.: *Papers in Structural and Transformational Linguistics*. D. Reidel, Dordrecht (1970)
- Johnson, H., Martin, J.: Unsupervised Learning of Morphology for English and Inuktitut. In: *Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada (2003)
- Kurimo, M., Creutz, M., Varjokallio, M.: Morpho Challenge Evaluation Using a Linguistic Gold Standard. In: *Proceedings of the CLEF 2007 Workshop*. Springer, Heidelberg (2008)
- Monson, C., Carbonell, J., Lavie, A., Levin, L.: ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis. In: *Computing and Historical Phonology: The Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, Prague, Czech Republic (2007)
- Snover, M.G.: An Unsupervised Knowledge Free Algorithm for the Learning of Morphology in Natural Languages. M.S. Thesis. Computer Science, Sever Institute of Technology, Washington University, Saint Louis, Missouri (2002)
- Stump, G.T.: *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press, Cambridge (2001)

SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval

Eneko Agirre¹, Oier Lopez de Lacalle¹, Bernardo Magnini², Arantxa Otegi¹,
German Rigau¹, and Piek Vossen³

¹ IXA NLP group, University of the Basque Country, Donostia, Basque Country

² ITC-IRST, Trento, Italy

³ Irion Technologies, Delftechpark 26, Delft, Netherlands

Abstract. This paper presents a first attempt of an application-driven evaluation exercise of WSD. We used a CLIR testbed from the Cross Lingual Evaluation Forum. The expansion, indexing and retrieval strategies were fixed by the organizers. The participants had to return both the topics and documents tagged with WordNet 1.6 word senses. The organization provided training data in the form of a pre-processed Semcor which could be readily used by participants. The task had two participants, and the organizer also provided an in-house WSD system for comparison. The results do not improve over the baseline, which is not surprising given the simplistic CLIR strategy used. Other than that the exercise was successful, and provides the foundation for more ambitious follow-up exercises where the participants would be able to build up on the WSD results already available.

1 Introduction

Since the start of Senseval, the evaluation of Word Sense Disambiguation (WSD) as a separate task is a mature field, with both lexical-sample and all-words tasks. In the first case the participants need to tag the occurrences of a few words, for which hand-tagged data has already been provided. In the all-words task all the occurrences of open-class words occurring in two or three documents (a few thousand words) need to be disambiguated.

The WSD community has long mentioned the necessity of evaluating WSD in applications, in order to check which WSD strategy is best suited for the application, and more important, to try to show that WSD can make a difference in applications. The successful use of WSD in Machine Translation has been the subject of some recent papers [4,3], but its contribution to Information Retrieval (IR) is yet to be shown. There have been with some limited experiments showing positive and negative evidence [14,7,10,11], with the positive evidence usually focusing on IR sub areas, such as CLIR [5,15] or Q&A [12]. [13] provides a nice overview of the applications of WSD and the issues involved.

With this proposal we want to make a first try in defining a task where WSD is evaluated with respect to an Information Retrieval and Cross-Lingual Information Retrieval (CLIR) exercise. From the WSD perspective, this task

will evaluate all-words WSD systems indirectly on a real task. From the CLIR perspective, this task will evaluate which WSD systems and strategies work best.

We are conscious that the number of possible configurations for such an exercise is very large (including sense inventory choice, using word sense induction instead of disambiguation, query expansion, WSD strategies, IR strategies, etc.), so this first edition focused on the following:

- The IR/CLIR system is fixed.
- The expansion / translation strategy is fixed.
- The participants can choose the best WSD strategy.
- The IR system is used as the upperbound for the CLIR systems.

We think that a focused evaluation where both WSD experts and IR experts use a common setting and shared resources might shed light to the intricacies in the interaction between WSD and IR strategies, and provide a fruitful ground for novel combinations and hopefully allow for breakthroughs in this complex area. We see this as the first of a series of exercises, and one outcome of this task should be that both WSD and CLIR communities discuss together future evaluation possibilities.

This task has been organized as a collaboration of SemEval¹ and the Cross-Language Evaluation Forum (CLEF²). The results were presented in both the SemEval-2007 and CLEF-2007 workshops, and a special track will be proposed for CLEF-2008, where CLIR systems will have the opportunity to use the annotated data produced as a result of the Semeval-2007 task. The task has a webpage with all the details at <http://ixa2.si.ehu.es/semeval-clir>.

This paper is organized as follows. Section 2 describes the task with all the details regarding datasets, expansion/translation, the IR/CLIR system used, and steps for participation. Section 3 presents the evaluation performed and the results obtained by the participants. Finally, Section 4 draws the conclusions and mention the future work.

2 Description of the Task

This is an application-driven task, where the application is a fixed CLIR system. Participants disambiguate text by assigning WordNet 1.6 synsets and the system will do the expansion to other languages, index the expanded documents and run the retrieval for all the languages in batch. The retrieval results are taken as the measure for fitness of the disambiguation. The modules and rules for the expansion and the retrieval will be exactly the same for all participants.

¹ <http://nlp.cs.swarthmore.edu/semeval/>

² <http://www.clef-campaign.org>

We proposed two specific subtasks:

1. Participants disambiguate the corpus, the corpus is expanded to synonyms/translations and we measure the effects on IR/CLIR. Topics³ are not processed.
2. Participants disambiguate the topics per language, we expand the queries to synonyms/translations and we measure the effects on IR/CLIR. Documents are not processed

The corpora and topics were obtained from the ad-hoc CLEF tasks. The supported languages in the topics are English and Spanish, but in order to limit the scope of the exercise we decided to only use English documents. The participants only had to disambiguate the English topics and documents. Note that most WSD systems only run on English text.

Due to these limitations, we had the following evaluation settings:

IR with WSD of documents, where the participants disambiguate the documents, the disambiguated documents are expanded to synonyms, and the original topics are used for querying. All documents and topics are in English.

IR with WSD of topics, where the participants disambiguate the topics, the disambiguated topics are expanded and used for querying the original documents. All documents and topics are in English.

CLIR with WSD of documents, where the participants disambiguate the documents, the disambiguated documents are translated, and the original topics in Spanish are used for querying. The documents are in English and the topics are in Spanish.

We decided to focus on CLIR for evaluation, given the difficulty of improving IR. The IR results are given as illustration, and as an upperbound of the CLIR task. This use of IR results as a reference for CLIR systems is customary in the CLIR community [8].

2.1 Datasets

The English CLEF data from years 2000-2005 comprises corpora from 'Los Angeles Times' (year 1994) and 'Glasgow Herald' (year 1995) amounting to 169,477 documents (579 MB of raw text, 4.8GB in the XML format provided to participants, see Section 2.3) and 300 topics in English and Spanish (the topics are human translations of each other). The relevance judgments were taken from CLEF. This might have the disadvantage of having been produced by pooling the results of CLEF participants, and might bias the results towards systems not using WSD, specially for monolingual English retrieval. We are considering the realization of a post-hoc analysis of the participants results in order to analyze the effect on the lack of pooling.

³ In IR topics are the short texts which are used by the systems to produce the queries. They usually provide extensive information about the text to be searched, which can be used both by the search engine and the human evaluators.

Due to the size of the document collection, we decided that the limited time available in the competition was too short to disambiguate the whole collection. We thus chose to take a sixth part of the corpus at random, comprising 29,375 documents (874MB in the XML format distributed to participants). Not all topics had relevant documents in this 17% sample, and therefore only 201 topics were effectively used for evaluation. All in all, we reused 21,797 relevance judgements that contained one of the documents in the 17% sample, from which 923 are positive⁴. For the future we would like to use the whole collection.

2.2 Expansion and Translation

For expansion and translation we used the publicly available Multilingual Central Repository (MCR) from the MEANING project [2]. The MCR follows the EuroWordNet design, and currently includes English, Spanish, Italian, Basque and Catalan wordnets tightly connected through the Interlingual Index (based on WordNet 1.6, but linked to all other WordNet versions).

We only expanded (translated) the senses returned by the WSD systems. That is, given a word like ‘car’, it will be expanded to ‘automobile’ or ‘railcar’ (and translated to ‘auto’ or ‘vagón’ respectively) depending on the sense in WN 1.6. If the systems returns more than one sense, we choose the sense with maximum weight. In case of ties, we expand (translate) all. The participants could thus implicitly affect the expansion results, for instance, when no sense could be selected for a target noun, the participants could either return nothing (or NOSENSE, which would be equivalent), or all senses with 0 score. In the first case no expansion would be performed, in the second all senses would be expanded, which is equivalent to full expansion. This fact will be mentioned again in Section 3.5.

Note that in all cases we never delete any of the words in the original text.

In addition to the expansion strategy used with the participants, we tested other expansion strategies as baselines:

noexp. No expansion, original text

fullexp. Expansion (translation in the case of English to Spanish expansion) to all synonyms of all senses

wsd50. Expansion to the best 50% senses as returned by the WSD system. This expansion was tried over the in-house WSD system of the organizer only.

2.3 IR/CLIR System

The retrieval engine is an adaptation of the TwentyOne search system [9] that was developed during the 90’s by the TNO research institute at Delft (The Netherlands) getting good results on IR and CLIR exercises in TREC [8]. It is now further developed by Irion technologies as a cross-lingual retrieval system [15]. For indexing, the TwentyOne system takes Noun Phrases as an input.

⁴ The overall figures are 125,556 relevance judgements for the 300 topics, from which 5700 are positive.

Noun Phases (NPs) are detected using a chunker and a word form with POS lexicon. Phrases outside the NPs are not indexed, as well as non-content words (determiners, prepositions, etc.) within the phrase.

The Irion TwentyOne system uses a two-stage retrieval process where relevant documents are first extracted using a vector space matching and secondly phrases are matched with specific queries. Likewise, the system is optimized for high-precision phrase retrieval with short queries (1 up to 5 words with a phrasal structure as well). The system can be stripped down to a basic vector space retrieval system with an tf.idf metrics that returns documents for topics up to a length of 30 words. The stripped-down version was used for this task to make the retrieval results compatible with the TREC/CLEF system.

The Irion system was also used for pre-processing. The CLEF corpus and topics were converted to the TwentyOne XML format, normalized, and named-entities and phrasal structured detected. Each of the target tokens was identified by a unique identifier.

2.4 Participation

The participants were provided with the following:

1. the document collection in Irion XML format
2. the topics in Irion XML format

In addition, the organizers also provided some of the widely used WSD features in a word-to-word fashion⁵ [1] in order to make participation easier. These features were available for both topics and documents as well as for all the words with frequency above 10 in SemCor 1.6 (which can be taken as the training data for supervised WSD systems). The Semcor data is publicly available⁶. For the rest of the data, participants had to sign an end user agreement.

The participants had to return the input files enriched with WordNet 1.6 sense tags in the required XML format:

1. for all the documents in the collection
2. for all the topics

Scripts to produce the desired output from word-to-word files and the input files were provided by organizers, as well as DTD's and software to check that the results were conformant to the respective DTD's.

3 Evaluation and Results

For each of the settings presented in Section 2 we present the results of the participants, as well as those of an in-house system presented by the organizers. Please

⁵ Each target word gets a file with all the occurrences, and each occurrence gets the occurrence identifier, the sense tag (if in training), and the list of features that apply to the occurrence.

⁶ <http://ixa2.si.ehu.es/semeval-clir/>

refer to the system description papers for a more complete description. We also provide some baselines and alternative expansion (translation) strategies. All systems are evaluated according to their Mean Average Precision ⁷ (MAP) as computed by the `trec_eval` software on the pre-existing CLEF relevance-assessments.

3.1 Participants

The two systems that registered sent the results on time.

PUTOP. They extend on McCarthy's predominant sense method to create an unsupervised method of word sense disambiguation that uses automatically derived topics using Latent Dirichlet Allocation. Using topic-specific synset similarity measures, they create predictions for each word in each document using only word frequency information. The disambiguation process took approx. 12 hours on a cluster of 48 machines (dual Xeons with 4GB of RAM). Note that contrary to the specifications, this team returned WordNet 2.1 senses, so we had to map automatically to 1.6 senses [6].

UNIBA. This team uses a knowledge-based WSD system that attempts to disambiguate all words in a text by exploiting WordNet relations. The main assumption is that a specific strategy for each Part-Of-Speech (POS) is better than a single strategy. Nouns are disambiguated basically using hypernymy links. Verbs are disambiguated according to the nouns surrounding them, and adjectives and adverbs use glosses.

ORGANIZERS. In addition to the regular participants, and out of the competition, the organizers run a regular supervised WSD system trained on Semcor. The system is based on a single k-NN classifier using the features described in [1] and made available at the task website (cf. Section 2.4).

In addition to those we also present some common IR/CLIR baselines, baseline WSD systems, and an alternative expansion:

noexp. A non-expansion IR/CLIR baseline of the documents or topics.

fullexp. A full-expansion IR/CLIR baseline of the documents or topics.

wsdrand. A WSD baseline system which chooses a sense at random. The usual expansion is applied.

1st. A WSD baseline system which returns the sense numbered as 1 in WordNet. The usual expansion is applied.

wsd50. The organizer's WSD system, where the 50% senses of the word ranking according to the WSD system are expanded. That is, instead of expanding the single best sense, it expands the best 50% senses.

3.2 IR Results

This section presents the results obtained by the participants and baselines in the two IR settings. The second and third columns of Table 1 present the results

⁷ http://en.wikipedia.org/wiki/Information_retrieval

Table 1. Retrieval results given as MAP. IRtops stands for English IR with topic expansion. IRdocs stands for English IR with document expansion. CLIR stands for CLIR results for translated documents.

	IRtops	IRdocs	CLIR
no expansion	0.3599	0.3599	0.1446
full expansion	0.1610	0.1410	0.2676
UNIBA	0.3030	0.1521	0.1373
PUTOP	0.3036	0.1482	0.1734
wsdrand	0.2673	0.1482	0.2617
1st sense	0.2862	0.1172	0.2637
ORGANIZERS	0.2886	0.1587	0.2664
wsd50	0.2651	0.1479	0.2640

when disambiguating the topics and the documents respectively. Non of the expansion techniques improves over the baseline (no expansion).

Note that due to the limitation of the search engine, long queries were truncated at 50 words, which might explain the very low results of the full expansion.

3.3 CLIR Results

The last column of Table 1 shows the CLIR results when expanding (translating) the disambiguated documents. None of the WSD systems attains the performance of full expansion, which would be the baseline CLIR system, but the WSD of the organizer gets close.

3.4 WSD Results

In addition to the IR and CLIR results we also provide the WSD performance of the participants on the Senseval 2 and 3 all-words task. The documents from those tasks were included alongside the CLEF documents, in the same formats, so they are treated as any other document. In order to evaluate, we had to

Table 2. English WSD results in the Senseval-2 and Senseval-3 all-words datasets

Senseval-2 all words			
	precision	recall	coverage
ORGANIZERS	0.584	0.577	93.61%
UNIBA	0.498	0.375	75.39%
PUTOP	0.388	0.240	61.92%
Senseval-3 all words			
	precision	recall	coverage
ORGANIZERS	0.591	0.566	95.76%
UNIBA	0.484	0.338	69.98%
PUTOP	0.334	0.186	55.68%

map automatically all WSD results to the respective WordNet version (using the mappings in [6] which are publicly available).

The results are presented in Table 2, where we can see that the best results are attained by the organizers WSD system.

3.5 Discussion

First of all, we would like to mention that the WSD and expansion strategy, which is very simplistic, degrades the IR performance. This was rather expected, as the IR experiments had an illustration goal, and are used for comparison with the CLIR experiments. In monolingual IR, expanding the topics is much less harmful than expanding the documents. Unfortunately the limitation to 50 words in the queries might have limited the expansion of the topics, which make the results rather unreliable. We plan to fix this for future evaluations.

Regarding CLIR results, even if none of the WSD systems were able to beat the full-expansion baseline, the organizers system was very close, which is quite encouraging due to the very simplistic expansion, indexing and retrieval strategies used.

In order to better interpret the results, Table 3 shows the amount of words after the expansion in each case. This data is very important in order to understand the behavior of each of the systems. Note that UNIBA returns 3 synsets at most, and therefore the *wsd50* strategy (select the 50% senses with best score) leaves a single synset, which is the same as taking the single best system (*wsdbest*). Regarding PUTOP, this system returned a single synset, and therefore the *wsd50* figures are the same as the *wsdbest* figures.

Comparing the amount of words for the two participant systems, we see that UNIBA has the least words, closely followed by PUTOP. The organizers WSD system gets far more expanded words. The explanation is that when the synsets returned by a WSD system all have 0 weights, the *wsdbest* expansion strategy expands them all. This was not explicit in the rules for participation, and might have affected the results.

A cross analysis of the result tables and the number of words is interesting. For instance, in the IR exercise, when we expand documents, the results in the third column of Table 1 show that the ranking for the non-informed baselines is the following: best for no expansion, second for random WSD, and third for full expansion. These results can be explained because of the amount of expansion: the more expansion the worst results. When more informed WSD is performed, documents with more expansion can get better results, and in fact the WSD system of the organizers is the second best result from all system and baselines, and has more words than the rest (with exception of *wsd50* and full expansion). Still, the no expansion baseline is far from the WSD results.

Regarding the CLIR result, the situation is inverted, with the best results for the most productive expansions (full expansion, random WSD and no expansion, in this order). For the more informed WSD methods, the best results are again for the organizers WSD system, which is very close to the full expansion baseline. Even if *wsd50* has more expanded words *wsdbest* is more effective. Note the very

Table 3. Number of words in the document collection after expansion for the WSD system and all baselines. wsdbest stands for the expansion strategy used with participants.

		English	Spanish
No WSD	noexp	9,900,818	9,900,818
	fullexp	93,551,450	58,491,767
UNIBA	wsdbest	19,436,374	17,226,104
	wsd50	19,436,374	17,226,104
PUTOP	wsdbest	20,101,627	16,591,485
	wsd50	20,101,627	16,591,485
Baseline	1st	24,842,800	20,261,081
WSD	wsdrand	24,904,717	19,137,981
ORG.	wsdbest	26,403,913	21,086,649
	wsd50	36,128,121	27,528,723

high results attained by random. These high results can be explained by the fact that many senses get the same translation, and thus for many words with few translation, the random translation might be valid. Still the wsdbest, 1st sense and wsd50 results get better results.

4 Conclusions and Future Work

This paper presents the results of a preliminary attempt of an application-driven evaluation exercise of WSD in CLIR. The expansion, indexing and retrieval strategies proved too simplistic, and none of the two participant systems and the organizers system were able to beat the full-expansion baseline. Due to efficiency reasons, the IRION system had some of its features turned off. Still the results are encouraging, as the organizers system was able to get very close to the full expansion strategy with much less expansion (translation).

All the resources built will be publicly available for further experimentations. We plan to propose a special track of CLEF-2008 where the participants will build on the resources (specially the WSD tagged corpora) in order to use more sophisticated CLIR techniques. We also plan to extend the WSD annotation to all words in the CLEF English document collection, and to contact the best performing systems of the SemEval all-words tasks to have better quality annotations.

Acknowledgements

We wish to thank CLEF for allowing us to use their data, and the CLEF coordinator, Carol Peters, for her help and collaboration. This work has been partially funded by the Spanish Education Ministry (project KNOW, TIN2006-15049-C03-01). Oier Lopez de Lacalle and Arantxa Otegi are supported by PhD grants from the Basque Government.

References

1. Agirre, E., de Lacalle, O.L., Martinez, D.: Exploring feature set combinations for WSD. In: Proc. of the SEPLN (2006)
2. Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P.: The MEANING Multilingual Central Repository. In: Proceedings of the 2.nd Global WordNet Conference, GWC 2004, pp. 23–30. Masaryk University, Brno (2004)
3. Carpuat, M., Wu, D.: Improving Statistical Machine Translation using Word Sense Disambiguation. In: Proc. of EMNLP-CoNLL, Prague (2007)
4. Chan, Y.S., Ng, H.T., Chiang, D.: Word Sense Disambiguation Improves Statistical Machine Translation. In: Proc. of ACL, Prague (2007)
5. Clough, P., Stevenson, M.: Cross-language information retrieval using EuroWordNet and word sense disambiguation. In: Proc. of ECIR, Sunderland (2004)
6. Daude, J., Padro, L., Rigau, G.: Mapping WordNets Using Structural Information. In: Proc. of ACL, Hong Kong (2000)
7. Gonzalo, J., Penas, A., Verdejo, F.: Lexical ambiguity and information retrieval revisited. In: Proc. of EMNLP, Maryland (1999)
8. Harman, D.: Beyond English. In: Voorhees, E.M., Harman, D. (eds.) TREC: Experiment and Evaluation in Information Retrieval, pp. 153–181. MIT press, Cambridge (2005)
9. Hiemstra, D., Kraaij, W.: Twenty-One in ad-hoc and CLIR. In: Voorhees, E.M., Harman, D.K. (eds.) Proc. of TREC-7, pp. 500–540. NIST Special Publication (1998)
10. Krovetz, B.: Homonymy and polysemy in information retrieval. In: Proc. of EACL, pp. 72–79, Madrid (1997)
11. Krovetz, B.: On the importance of word sense disambiguation for information retrieval. In: Proc. of LREC Workshop on Creating and Using Semantics for Information Retrieval and Filtering, Las Palmas (2002)
12. Pasca, M., Harabagiu, S.: High performance question answering. In: Proc. of ACM SIGIR, New Orleans (2001)
13. Resnik, P.: Word sense disambiguation in nlp applications. In: Agirre, E., Edmonds, P. (eds.) Word Sense Disambiguation: Algorithms and Applications. Springer, Heidelberg (2006)
14. Voorhees, E.M.: Natural language processing and information retrieval. In: Pazienza, M.T. (ed.) Information Extraction: Towards Scalable, Adaptable Systems. Springer, Heidelberg (1999)
15. Vossen, P., Rigau, G., Alegria, I., Agirre, E., Farwell, D., Fuentes, M.: Meaningful results for Information Retrieval in the MEANING project. In: Proc. of the 3rd Global Wordnet Conference, pp. 22–26 (2006)

Author Index

- Abdou, Samir 196
 Adamek, Tomasz 530
 Adda, Gilles 414
 Adiwbowo, Septian 332
 Adriani, Mirna 127, 332, 742, 838
 Agirre, Eneko 908
 Ahn, David 344
 Airio, Eija 530
 Al-Jumaily, Harith 352
 Alzghool, Muath 695
 Amaral, Carlos 364
 Andogah, Geoffrey 794
 Argaw, Atelach Alemu 119
 Ayache, Christelle 200, 249

 Baerisch, Stefan 160
 Balahur, Alexandra 336
 Balahur-Dobrescu, Alexandra 395
 Bandyopadhyay, Sivaaji 88
 Banerjee, Pratyush 95
 Batista, Fernando 356
 Benajiba, Yassine 324
 Benczúr, András 72, 445, 518
 Bergler, Sabine 562
 Bernhard, Delphine 873
 Berrocal, José L. Alonso 143, 732
 Bhattacharyya, Pushpak 111, 657
 Bilinski, Eric 414
 Bíró, István 72, 518
 Bordag, Stefan 881
 Boughanem, Mohand 665
 Bouma, Gosse 257, 794
 Bowden, Mitchell 273
 Brailsford, Tim 64
 Brendel, Mátyás 445, 518
 Buscaldi, Davide 324, 815

 Cabral, Luís Miguel 261
 Caicedo, Juan C. 615
 Caputo, Barbara 577
 Carbonell, Jaime 900
 Cardoso, Nuno 802
 Cassan, Adán 364
 Ceaşu, Alexandru 284
 Česka, Pavel 33

 Chang, Yih-Chen 504
 Charniak, Eugene 687
 Chaves, Marcirio 802
 Chen, Hsin-Hsi 504
 Chevallet, Jean-Pierre 585, 631
 Chinnakotla, Manoj Kumar 111
 Clinchant, Stephane 182, 569
 Clough, Paul 433
 Coheur, Luísa 356
 Comas, Pere R. 249, 424
 Costa, Luís Fernando 261
 Cotelea, Diana 336
 Creutz, Mathias 864
 Cristea, Dan 336
 Cruz, David 802
 Csalogány, Károly 72, 518
 Csurka, Gabriela 569

 d'Silva, Thomas 273
 Damani, Om P. 111
 Dandapat, Sandipan 95
 Daróczy, Bálint 445, 518
 de Matos, David Martins 356
 Denicia, Claudia 328
 de Pablo-Sánchez, César 352
 de Rijke, Maarten 344, 725, 737
 Desai, Bipin C. 308, 657
 Deselaers, Thomas 445, 472, 492
 Deserno, Thomas M. 472, 637
 Díaz-Galiano, M.C. 512, 601, 719
 Di Nunzio, Giorgio M. 13, 745
 Dolamic, Ljiljana 37, 196
 Dornescu, Iustin 336
 Drăghici, Iuliana 336

 Eibl, Maximilian 174
 Ekbal, Asif 88
 El Demerdash, Osama 562
 Escalante Balderas, Hugo Jair 445, 546

 Fautsch, Claire 196
 Ferrández, Óscar 377
 Ferrés, Daniel 830, 834
 Ferro, Nicola 13, 745

- Figueira, Helena 364
 Figuerola, Carlos G. 143, 732
 Firgantoro, Tri 742
 Forascu, Corina 200
 Forner, Pamela 200
- Galibert, Olivier 414
 García-Cumbreras, M.A. 137, 381, 512, 601, 719, 823
 García-Vega, Manuel 823
 Gass, Tobias 492
 Gaussier, Eric 585
 Gevers, Theo 445
 Gey, Fredric 745
 Giampiccolo, Danilo 200
 Glöckner, Ingo 269, 372
 Goñi-Menoyo, José Miguel 156, 597, 786
 Gobeill, Julien 649
 Godhavarthy, Srinivasa Rao 88
 González-Cristóbal, José Carlos 156, 500, 593, 597, 786
 González-Ledesma, Ana 352
 Gonzalez, Fabio A. 615
 Grubinger, Michael 433
 Guillén, Rocio 781
 Guillena, Rafael Muñoz 522
 Güld, Mark O. 637
 Gupta, Mayank 95
- Haddad, Chedid 308
 Hanbury, Allan 433, 445
 Haque, Rejwanul 88
 Hartrumpf, Sven 269, 773
 Hayurani, Herika 127
 Hendriansyah, Okky 742
 Hernández, Carlos 546
 Hernández, Gustavo 328
 Hernández Gracidas, Carlos Arturo 445
 Herrera, Jesús 200
 Hersh, William 472, 623
 Heuwing, Ben 134, 850
 Hoffmannová, Petra 674
 Hofmann, Katja 344
 Hoi, Steven C.H. 445, 538
- Iftene, Adrian 336, 395
 Inkpen, Diana 695
 Ion, Radu 284
 Ircing, Pavel 712
- Jagarlamudi, Jagadeesh 80
 Järvelin, Anni 530
 Jijkoun, Valentin 200, 344, 725, 737
 Jones, Gareth J.F. 530, 674, 703
 Juárez, Antonio 328
- Kalpathy-Cramer, Jayashree 472, 623
 Khalid, Mahboob Alam 344
 Kim, Eugene 472
 Kloosterman, Geert 257
 Kölle, Ralph 850
 Kosseim, Leila 562
 Kumaran, A. 80
 Kurimo, Mikko 864
 Kürsten, Jens 174
- Laaksonen, Jorma 445
 Lamel, Lori 249
 Lana-Serrano, Sara 156, 500, 593, 597, 786
 Larson, Ray R. 188, 745, 811
 Lavie, Alon 900
 Le, Thi Hoang Diem 631
 Lease, Matthew 687
 Leveling, Johannes 269, 773
 Levin, Lori 900
 Li, Mingjing 445
 Li, Zhisheng 842
 Lim, Joo Hwee 631
 Llopis, Fernando 45, 522
 López, Aurelio 546
 Lopez de Lacalle, Oier 908
- Ma, Wei-Ying 842
 Magalhães, João 856
 Magnini, Bernardo 908
 Maisonnasse, Loic 585
 Majumder, Prasenjit 49
 Mamede, Nuno J. 356
 Mandal, Debasis 95
 Mandl, Thomas 13, 134, 745, 850
 Marín Castro, Heidy Marisol 445
 Marín, Heidy 546
 Martín-Valdivia, M.T. 512, 601, 719
 Martínez, Paloma 352
 Martínez-Fernández, José Luis 352, 500
 Martínez-Santiago, F. 137, 381
 Martins, André 364
 Mendes, Afonso 364
 Mendes, Ana 356

- Mendes, Pedro 364
 Micol, Daniel 377
 Mitra, Mandar 49
 Moldovan, Dan 273
 Mondal, Tapabrata 88
 Monson, Christian 900
 Montejó-Ráez, A. 137, 512, 601
 Montes, Manuel 328, 546
 Montes-y-Gómez, Manuel 391
 Morales, Eduardo 546
 Moreno-Sandoval, Antonio 352
 Moruz, Alex 336
 Mostefa, Djamel 249
 Müller, Henning 433, 472, 649
 Muñoz, Rafael 377
 Mur, Jori 257

 Nasikhin 838
 Naskar, Sudip Kumar 88
 Navarro, Sergio 522
 Neumann, Günter 292, 387, 410
 Névél, Aurélie 641
 Ney, Hermann 445, 492
 Noguera, Elisa 45, 522

 Oakes, Michael P. 148
 Oard, Douglas W. 674
 Olteanu, Marian 273
 Orabona, Francesco 577
 Orăsan, Constantin 300
 Osenova, Petya 200
 Otegi, Arantxa 908
 Overell, Simon 856

 Pal, Dipasree 49
 Palomar, Manuel 377
 Peñas, Anselmo 200, 237, 404
 Pecina, Pavel 33, 674
 Perea-Ortega, José M. 381, 823
 Peters, Carol 1, 13
 Petras, Vivien 160
 Pinel-Sauvagnat, Karen 665
 Pingali, Prasad 103
 Pinto, Cláudia 364
 Pistol, Ionuț 336
 Psutka, Josef 712
 Pușcașu, Georgiana 300

 Quaresma, Paulo 316

 Rahman, M.M. 657
 Ranadive, Sagar 111
 Renders, Jean-Michel 182, 569
 Ribeiro, Ricardo 356
 Rigau, German 908
 Rocha, Paulo 200
 Rodríguez, Horacio 830, 834
 Rodrigo, Álvaro 237, 404
 Romero, Eduardo 615
 Rosset, Sophie 249, 414
 Rosso, Paolo 324, 815
 Ruch, Patrick 649
 Rüger, Stefan 856
 Rui, Xiaoguang 445
 Ruiz, Miguel E. 641

 Sacaleanu, Bogdan 200, 292
 Saias, José 316
 Samy, Doaa 352
 Sanchis, Emilio 324
 Sanderson, Mark 745
 Santos, Diana 261, 745
 Sari, Syandra 127
 Sarkar, Sudeshna 95
 Savoy, Jacques 37, 196
 Schönhofen, Péter 72
 Schuldt, Heiko 607
 Sebe, Nicu 445
 Siklósi, Dávid 518
 Silva, Mário J. 802
 Smeaton, Alan F. 530
 Sormunen, Eero 530
 Springmann, Michael 607
 Spurk, Christian 292
 Ștefănescu, Dan 284
 Stempfhuber, Maximillian 160
 Stöttinger, Julian 445
 Sucar, Enrique 546
 Surdeanu, Mihai 424
 Suriyentrakorn, Pasin 273
 Sutcliffe, Richard 200

 Téllez-Valero, Alberto 328, 391
 Tiedemann, Jörg 257
 Tjong Kim Sang, Erik 344
 Tomlinson, Stephen 57
 Tommasi, Tatiana 577
 Torjmen, Mouna 665
 Trandabăt, Diana 336
 Truran, Mark 64

- Tufiş, Dan 284
 Tune, Kula Kekeba 103
 Turmo, Jordi 249, 424
 Tushabe, Florence 554

 Ureña-López, L.A. 381, 512, 601,
 719, 823

 van der Plas, Lonneke 257
 van Rantwijk, Joris 344
 van Noord, Gertjan 257
 Varjokallio, Matti 864
 Varma, Vasudeva 103
 Vavruška, Jan 712
 Verdejo, Felisa 237, 404
 Vidal, Daniel 364
 Viitaniemi, Ville 445
 Vilares, Jesús 148
 Vilares, Manuel 148
 Villaseñor, Luis 328, 546
 Villaseñor-Pineda, Luis 391
 Villatoro, Esaú 328

 Villena-Román, Julio 156, 500, 593,
 597, 786
 Vossen, Piek 908

 Wang, Chong 842
 Wang, Rui 387, 410
 Wang, Xufa 842
 Weyand, Tobias 492
 Wilhelm, Thomas 174
 Wilkins, Peter 530
 Wilkinson, Michael. H.F. 554
 Womser-Hacker, Christa 745, 850
 Wu, Lei 445

 Xie, Xing 745, 842

 Zazo, Angel F. 143
 Zazo Rodríguez, Angel F. 732
 Zeman, Daniel 892
 Zhang, Ke 703
 Zhang, Ying 674, 703
 Zhou, Dong 64
 Zhou, Xin 649