Cross-Language System Evaluation Campaign

CLEF 2008

Book of Abstracts

Edited by
Francesca Borri, Alessandro Nardi, Carol Peters
ISTI-CNR, Pisa, Italy

ISSN: 1818-8044
ISBN: 2-912335-43-4

CONTENTS

ImageCLEFWiki........75

Multilingual Web Track (WebCLEF)........80

What Happened in CLEF 2008

Carol Peters
ISTI-CNR, Area di Ricerca, Italy
carol.peters@isti.cnr.it

This volume contains a set of abstracts that summarize the experiments conducted in CLEF 2008 - the ninth information system evaluation campaign organized by the Cross-Language Evaluation Forum[1]. It has been prepared for distribution at the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark, together with an electronic version of the complete CLEF 2008 Working Notes. The Working Notes provide a first description of the various experiments made by this year's participants, preliminary analyses of results by the track coordinators, and appendices containing run statistics and overview graphs for some of the tracks/tasks. They are also available on-line on the CLEF website www.clef-campaign.org. The main features of the 2008 campaign are briefly outlined here below. More details can be found in the Track Overviews in this volume and in the complete Working Notes.

CLEF 2008 Tracks

CLEF 2008 offered seven tracks designed to evaluate the performance of systems for:

- multilingual textual document retrieval (Ad Hoc)
- mono- and cross-language information retrieval on structured scientific data (Domain-Specific)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval in image collections (ImageCLEF)
- multilingual retrieval of Web documents (WebCLEF)
- cross-language geographical information retrieval (GeoCLEF)

Two new tracks were offered as pilot tasks:

- cross-language video retrieval (VideoCLEF)
- multilingual information filtering (INFILE@CLEF)
- In addition, Morpho Challenge 2008 was organized in collaboration with CLEF[2]

Test Collections

A number of document collections were used in CLEF 2008 to build the test collections, including:

- CLEF multilingual corpus of more than 3 million news documents in 14 European languages[3]
- Hamshahri Persian newspaper corpus, 2000-2006
- About 3 million library catalog records in English, French, German, derived from The European Library
- GIRT-4 English/German social science database, the Russian ISISS collection for sociology and economics and Cambridge Sociological Abstracts
- The ImageCLEF track used collections for both general photographic and medical image retrieval:
- IAPR TC-12 photo database; INEX Wikipedia image collection
- ARRS Goldminer database of radiographs; IRMA collection for medical image annotation
- Dutch and English documentary television programs provided by Sound & Vision, The Netherlands
- Agence France Press (AFP) comparable newswire stories in Arabic, French and English

Participation

A total of 100 groups submitted runs in CLEF 2008, a big increase on the 81 groups of CLEF 2007: 69 from Europe, 12 from N.America; 15 from Asia, 3 from S.America and 1 from Africa. The breakdown of participation of groups per track is as follows: Ad Hoc 26; Domain-Specific 6; iCLEF 6; QAatCLEF 29; ImageCLEF 42; WebCLEF 3; GeoCLEF 11; VideoCLEF 5; INFILE 1; Morpho Challenge 6. A list of groups and indications of the tracks in which they participated is given in the Appendix to this volume.

I should like to conclude by thanking everyone who has contributed to the success of CLEF 2008: the Steering Committee, the Track Coordinators, collaborating institutions and individuals, data providers, and last but certainly not least all the participants, who I hope have found CLEF to be a valuable and rewarding experience. Let me end by wishing everyone an interesting, worthwhile and above all enjoyable Workshop!

[1] CLEF is an activity of the EU 7FP TrebleCLEF Coordination Action, see http://www.trebleclef.eu

[2] Morpho Challenge is part of the EU Network of Excellence Pascal: http://www.cis.hut.fi/morphochallenge2008/

[3] This corpus currently contains news documents for the same time period (1994-95) in ten languages: Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, and Swedish, and for 2000-2002 in Basque, Czech, Bulgarian English and Hungarian.

Multilingual Textual Document Retrieval (Ad Hoc)

CLEF 2008: Ad Hoc Track Overview

Eneko Agirre[1], Giorgio M. Di Nunzio[2], Nicola Ferro[2], Thomas Mandl[3], and Carol Peters[4]
[1]Computer Science Department, University of the Basque Country, Spain
[2]Department of Information Engineering, University of Padua, Italy
[3]Information Science, University of Hildesheim – Germany
[4]ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
e.agirre@ehu.es, {dinunzio, ferro}@dei.unipd.it, mandl@uni-hildesheim.de, carol.peters@isti.cnr.it

The aim of the ad hoc track is to promote the development of monolingual and cross-language textual document retrieval systems. From 2000-2007, the track used exclusively collections of European newspaper and news agency documents[1]. This year, we widened our scope by introducing very different document collections, a non-European target language, and an information retrieval (IR) task designed to attract participation from groups interested in natural language processing (NLP). The track was thus structured in three distinct streams:

- TEL@CLEF
- Persian@CLEF
- Robust WSD

A total of 24 groups from 14 different countries submitted results for one or more of these tasks - a slight increase on the 22 participants of last year.

The first task offered monolingual and cross-language search on library catalog records. It was organized in collaboration with The European Library (TEL)[2] and used three collections from the catalogs of the British Library, the Bibliothèque Nationale de France and the Austrian National Library. The underlying aim was to identify the most effective retrieval technologies for searching this type of very sparse multilingual data. In fact, the collections contained records in many languages in addition to English, French or German. The task presumed a user with a working knowledge of these three languages who wants to find documents that can be useful for them in one of the three target catalogs. Records in other languages were counted irrelevant. This was a challenging task but proved popular; participants tried various strategies to handle the multilinguality of the catalogs. The fact that the best results were not always obtained by experienced CLEF participants shows that the traditional approaches used for newspaper document retrieval are not necessarily the most effective for this type of data. We are still analysing the results to see what can be learned; participants will be encouraged to do in-depth analyses and the task will certainly be offered again in CLEF 2009.

The Persian@CLEF activity was coordinated in collaboration with the Database Research Group (DBRG) of Tehran University. It was the first time that CLEF offered a non-European language target collection. We chose Persian for several reasons: its challenging script (a modified version of the Arabic alphabet with elision of short vowels) written from right to left; its complex morphology (extensive use of suffixes and compounding); its political and cultural importance. The task used the Hamshahri corpus of 1996-2002 newspapers as the target collection. Monolingual and cross-language (English to Persian) tasks were offered. As was to be expected, many of the eight participants focused their attention on problems of stemming. Only three submitted cross-language runs. The results of the best groups were in line with previous CLEF ad hoc experiments.

The robust task ran for the third time at CLEF 2008. This year it used English test data from previous campaigns but, in addition to the original documents and topics, the organizers provided word sense disambiguated (WSD) documents and topics. Both monolingual and bilingual experiments (topics in Spanish) were activated. The results for the 8 participating groups were mixed: while some top scoring groups did manage to improve the results using WSD information in both monolingual and bilingual settings, and the best monolingual robustness (GMAP) score was for a WSD run, the best scores for the rest came from systems which did not use WSD information. Given the relatively short time that the participants had to try effective ways of using the word sense information we think that these results are very positive, and a subsequent evaluation exercise would be needed for participants to further develop their systems.

[1] Over the years, test collections for mono- and cross-language system evaluation in 13 European languages have been created.

[2] http://www.theeuropeanlibrary.org/

TEL@CLEF

Logistic Regression for Metadata: Cheshire takes on Adhoc-TEL

Ray R. Larson
University of California, Berkeley, School of Information
ray@sims.berkeley.edu

In this paper we will briefly describe the approaches taken by the Berkeley Cheshire Group for the Adhoc-TEL 2008 tasks (Mono and Bilingual retrieval). All of the submitted runs were automatic without manual intervention in the queries (or translations). We submitted six Monolingual runs (two German, two English, and two French) and nine Bilingual runs (each of the three main languages to both of the other main languages (German, English and French). In addition we submitted three runs from Spanish translations of the topics to the three main languages.

Since the Adhoc-TEL task is new for this year, we took the approach of using methods that have performed fairly well in other tasks. In particular, the approach this year used probabilistic text retrieval based on logistic regression and incorporating blind relevance feedback for all of the runs. All translation for bilingual tasks was performed using the LEC Power Translator PC-based MT system. This approach seems to be a good fit for the limited TEL records, since the overall results show Cheshire runs in the top five submitted runs for all languages and tasks except for Monolingual German.

CACAO Project at the TEL@CLEF 2008 Task

Alessio Bosca and Luca Dini
Celi s.r.l - 10131 Torino - C. Moncalieri, 21
{alessio.bosca, dini}@celi.it

The paper describes the participation of the CACAO project consortium to the TEL@CLEF 2008 task targeted at retrieving relevant items from collections of library catalogues. CACAO project proposes the development of an infrastructure for multilingual access to digital content, including an information retrieval system able to search for books and texts in all the available languages. For each monolingual and bilingual subtask two different experiments have been conducted, one involving additional query expansion and one not.

Some of the textual information contained in the collections metadata (dc:subject, dc:title and dc:dexcription) has been lemmatised using the XIP incremental parser from XEROX and all the data has been then indexed using the Lucene open source engine.

By means of lexical semantics technologies a corpus based word space model has been created for each of the TEL@CLEF collections; these word space resources have been used by the CACAO system as a means to disambiguate the candidate translations and for query expansion purposes.

The approach adopted by CACAO system for dealing with user queries is based on the free keywords search; therefore while the title field of TEL topics already fitted this model, the description field has been processed in order to extract a set of relevant keywords from the sentence. For this purpose a simple keyword extractor module has been used for each of the main languages present in the corpus (English, French and German).

Each description sentence has been analysed in order to extract two different kinds of information, one representing the content type of the items to be retrieved (as novels, poetry or photo collections) and the other conveying additional detail on user interests.

The translation process exploited internal resources (inter-lingual indexes or bilingual dictionaries) and online dictionaries as Ergane; the so-obtained translation candidates have been disambiguated using the corpus based semantic vectors. Experiments involving query expansion enriched the keywords groups exploiting the corpus based semantic vectors.

Results evidenced a poor initial performance of the system, however since the project started few months ago they can constitute a valuable baseline in order to measure the future advancement of the system.

Technical University of Lisbon CLEF 2008 Submission: TEL@CLEF Monolingual Task

Jorge Machado, Bruno Martins and José Borbinha
Departmento de Engenharia Informática, Technical University of Lisbon, Portugal

We describe our participation in the TEL@CLEF monolingual tasks of the CLEF 2008 ad-hoc track, where we measured the retrieval performance of the IR service that is currently under development as part of the DIGMAP project (www.digmap.eu). DIGMAP's IR service is mostly based on Lucene, together with extensions for using query expansion and multinomial language modelling. In our runs, we experimented combinations of Rocchio query expansion, Lucene's off-the-shelf ranking scheme and the ranking scheme based on multinomial language modelling. The Lucene extensions that we used were LucQE (http://lucene-qe.sourceforge.net/) for query expansion and LM-Lucene (http://ilps.science.uva.nl/Resources/) for language modelling. Results show that query expansion and multinomial language modelling both result in

increased performance. Our best official run used the language modelling extension together with stemming, with no query expansion. This run achieved MAP scores of 0.3623, 0.2341 and 0.2298, respectively for the BL (English), Bnf (French) and ONG (German) collections.

UFRGS@CLEF2008: Using Association Rules for Cross-Language Information Retrieval

André Pinto Geraldo and Viviane Moreira Orengo
Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
[apgeraldo,vmorengo]@inf.ufrgs.br

This paper reports on monolingual and bilingual ad-hoc information retrieval experiments that we have performed for the TEL task at CLEF2008. Our aim was to use algorithms for mining association rules (ARs) to map concepts between languages on a Cross-Language Information Retrieval (CLIR) scenario. These algorithms are widely used for data mining purposes. A common example is market-basket data, i.e. the items that a customer buys at one transaction. For such data, an association rule would state, for example, that "90% of customers that purchase bread also purchase milk".

Our proposal is to map the problem of finding ARs between items in a market-basket scenario to the problem of finding cross-linguistic equivalents between a pair of languages on a parallel corpus. This approach is based on co-occurrences and works under the assumption that cross-linguistic equivalents would have a significant number of co-occurrences over a parallel corpus.

The proposed approach to use algorithms for mining ARs for CLIR can be divided into five phases: (i) pre-processing, (ii) mining ARs, (iii) rule filtering, (iv) query translation, and (v) query execution. Each term in the original query is replaced by all possible translations that remain after the filtering process.

We worked on the English TEL collection, which contains catalogue data from the British Library. Our bilingual runs used Spanish topics to query English texts. We also used the Porter stemmer for English and its Spanish version to remove suffixes. We used the Apriori algorithm for mining ARs and Zettair was the search engine chosen.

The results of the experiments show that the performance of our approach is not statistically different from the monolingual baseline in terms of mean average precision. This is an indication that association rules can be effectively used to map concepts between languages.

We have also tested a modification to BM25 that aims at increasing the weight of rare terms. The results show that this modified version achieved better performance. The improvements were considered to be statistically significant in terms of MAP on our monolingual runs.

WikiTranslate: Query Translation for Cross-lingual Information Retrieval Using only Wikipedia

D. Nguyen, A.Overwijk, C.Hauff, R.B. Trieschnigg, D. Hiemstra and F.M.G. de Jong
Twente University
{dong.p.ng, arnold.overwijk}@gmail.com, {c.hauff, trieschn,f.m.g.dejong}@ewi.utwente.nl, hiemstra@cs.utwente.nl

This paper presents WikiTranslate, a system which performs query translation for cross-lingual information retrieval (CLIR) using only Wikipedia to obtain translations.

We treat Wikipedia articles as representations of concepts (i.e. units of knowledge). WikiTranslate maps the query to Wikipedia concepts. Through the cross-lingual links translations of the concepts into language-specific terms are retrieved. The system makes use of the unique features of Wikipedia (e.g. the text, title, cross-lingual links, internal links and redirect pages).

The first step maps the query to Wikipedia concepts. First the most relevant concepts to the query are extracted after a search with the whole query (step 1a) in the Wikipedia articles. Then a search on every term of the query is performed (step 1b). This is done in two different ways. Using the internal links from the concepts retrieved with step 1a and using the text and title of the Wikipedia articles.

The second step creates the translated query. First we add articles that redirect to the found Wikipedia concepts to include synonyms and spelling variants. Furthermore articles retrieved with step 1a are given more weight. Finally, the final query is created using the found concepts.

WikiTranslate is evaluated by searching with topics in Dutch, French and Spanish in an English data collection. The systems achieved a performance of 67% compared to the monolingual baseline.

Cross-language Information Retrieval using Explicit Semantic Analysis

Philipp Sorg and Philipp Cimiano
Institute AIFB, Univeristät Karlsruhe (TH)
{sorg,cimiano}@aifb.uni-karlsruhe.de

We have participated on the Monolingual and Bilingual Ad-Hoc Retrieval Tasks, using a novel extension of the by now well-known Explicit Semantic Analysis (ESA) approach. We call this extension Cross-Language Explicit Semantic Analysis (CL-ESA) as it allows to apply ESA in a cross-lingual information retrieval setting. In essence, ESA represents documents as real-valued vectors in the space of Wikipedia articles, using the tf.idf measure to capture how "important" a Wikipedia article is for a specific document. The interesting property of ESA is that arbitrary documents can be represented as a vector with respect to the Wikipedia article space. Hereby, Wikipedia's articles are thus used as universal categories with respect to which arbitrary texts can be indexed. ESA thus essentially replaces the BOW model, but keeping its traditional operations.

In our cross-lingual extension of ESA, the cross-language links of Wikipedia are used in order to map the ESA vectors between different languages, thus allowing retrieval across languages. This requires to detect the language of each document in the collection in order to index it with respect to the correct Wikipedia database. For the cross-lingual retrieval we used the cosine similarity measure on the mapped vectors to compute a ranking of documents for queries.

Currently, our implementation supports the three languages English, German and French but could be extended to other languages by indexing Wikipedia databases in other languages.

The main objectives of our experiments was to discover if CL-ESA performs well in a cross-lingual retrieval setting and that it could therefore be used to build new CLIR systems competing with current state-of-the-art approaches.

As resources we rely on standard stemming techniques (Snowball stemmer), on Lucene as index and retrieval engine in order to index Wikipedia articles as well as on the Wikipedia database dump to index items in the document collections with respect to the Wikipedia article space.

Our results are far behind the ones of other systems, but we are confident that there is a large margin to improve the system. Methods to refine the ESA vector have been shown to improve IR results substantially. This can also be applied to the cross-lingual case.

CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval

Jens Kürsten, Thomas Wilhelm and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media, 09107 Chemnitz, Germany
[jens.kuersten j thomas.wilhelm j maximilian.eibl] at cs.tu-chemnitz.de

This article describes our first participation at the Ad-Hoc track. We used the Xtrieval framework for the preparation and execution of the experiments. Our main goal was to address the multilingual content of the provided collection by applying language detection in the indexing phase. Unfortunately, we had to omit this idea because of a mistake in the indexing procedure that was recognized for the first time on the day of the submission deadline. We regard our experiments as on-line or live experiments since the preparation of all results including indexing and retrieval took us less than 4 hours in total.

This year, we submitted 18 experiments, whereof only 4 were pure monolingual runs. In all our experiments we applied a standard top-k pseudo-relevance feedback algorithm. The translation of the topics for the multilingual experiments was realized with a plug-in to access the Google AJAX language API. The performance of our monolingual experiments was slightly below the average for the German and French collection and in the top 5 for the English collection. Our bilingual experiments performed very well (at least in the top 3) for all target collections. Generally speaking, the performance of our experiments amazingly exceeded our expectations, especially when having in mind the problems we faced only a few hours before the submission deadline. Finally, we would like to state that the strong performance of our cross-lingual experiments is most likely to be credited to the quality of the translation of topics.

XRCE's Participation to CLEF2008 - Ad-hoc Track

S. Clinchant and J.M. Renders
Xerox Research Centre Europe, 6 ch. de Maupertuis, 38240 Meylan, France
FirstName.LastName@xrce.xerox.com

Our participation to CLEF2008 (Ad-Hoc Track, TEL Subtask) was an opportunity to develop and assess methods that tackle multilinguality in a principled - while rather simple - way. It was also an opportunity to demonstrate the effectiveness of the dictionary adaptation method we designed last year in the case of the domain-specific track.

Our goal was to get a single retrieval model and index for all the languages of one specific collection. However, this approach required to give weights to each language to merge dictionaries at retrieval time. While assigning such weights requires prior knowledge about the collections, the dictionary adaptation mechanism provides a partial solution to this problem, adapting weights to each query.

Unfortunately, the accumulation of some mistakes rendered our official runs relatively inefficient. In particular, a misunderstanding of the ``bilingual task" definition led us not to index a significant part of the collections. In this note, we present the reasons of these mistakes and report how we partly corrected some of them in a set of extra unofficial runs whose performances are among the best ones; they demonstrate that dictionary adaptation is effective for the TEL task and corpora. This set of extra experiments is based on a simplifying assumption that considers all bilingual tasks as really bilingual, with one source language and one unique target language (the official language of the target collection). Further work will require re-processing the collections to keep the documents we have not indexed.

We will also need to come back to a true multilingual setting by solving the issue of weighting differently the basic bilingual lexicons and monolingual thesauri.

Persian@CLEF

Cross Language Experiments at Persian@CLEF 2008

Abolfazl AleAhmad[1], Ehsan Kamalloo[1], Arash Zareh[1], Masoud Rahgozar[1] and Farhad Oroumchian
[1]Database Research Group, School of Electrical and Computer Engineering
{a.aleahmad, e.kamalloo, a.zareh}@ece.ut.ac.ir,
rahgozar@ut.ac.ir
Department of Computer Science, University of Wollongong in Dubai
oroumchian@acm.org

In this study we will discuss our cross language text retrieval (CLIR) experiments of Persian ad hoc track at CLEF 2008. Two teams from University of Tehran were involved in cross language text retrieval part of the track using two different CLIR approaches that are query translation and document translation. For query translation we used a method named Combinatorial Translation Probability (CTP) calculation for estimation of translation probabilities. In the document translation part we used the Shiraz machine translation system for translation of documents into English. Then we create a Hybrid CLIR system by score-based merging of the two retrieval system results. In addition, we investigated N-grams and a light stemmer in our monolingual experiments.

Using Part of Speech Tagging in Persian Information Retrieval

Reza Karimpour, Amineh Ghorbani, Azadeh Pishdad, Mitra Mohtarami, Abolfazl AleAhmad, Hadi Amiri and Farhad Oroumchian
Database Research Group, University of Tehran
rezaka@gmail.com, m.mohtarami@yahoo.com, {a.ghorbany,a.pishdad,a.aleahmad,h.amiri}@ece.ut.ac.ir
foroumchian@acm.org

The text retrieval methods may benefit from natural language constructs to boost their results by achieving higher precision/recall rates. In this attempt, among many natural language features, we have used part of speech attributes of terms as extra information about document and query terms and have evaluated the impact of such information on the performance of the retrieval algorithms. Also the effect of stemming was investigated as a complement to this research.

In this research Bijankhan manually tagged collection of Persian words was used to train TnT part of speech tagger and thereafter Hamshahri Persian Corpus was tagged by TnT. Also stemmed versions of Hamshahri corpus both non-tagged and tagged were developed by a simple grammatical based tool called PERSTEM. Since Indri (part of Lemur project) is a language modelling based tool with weighting support, the experiments were conducted using the indri retrieval system along with training data that comes with Hamshahri corpus. According to information importance of part of speech of the words, query terms were weighted. Also, different weighting schemas as well as the omission of less important tags were experimented. After analyzing the impact of different tags, eventually we find out that noun, verb, adjective, and adverb are the most important POS Tags in Persian retrieval.

In contrast with experiments conducted by other groups in University of Tehran on Hamshahri corpus, the results we obtained indicate that the Persian retrieval benefits from stemming. Our findings suggest that part of speech tags may have small influence on effectiveness of the retrieved results. However, when this information is combined with stemming, accuracy of the retrievals is improved considerably. In order to further benchmark our system we decided to participate in CELF 2008 monolingual Persian ad hoc track. Our system named Tehran-NLP could not make it to the top five because of some technical problems.

Investigation on Application of Local Cluster Analysis and Part of Speech Tagging on Persian Text

Amir Hossein Jadidinejad[1], Mitra Mohtarami[2] and Hadi Amiri[2]
[1]Computer Engineering Department, Islamic Azad University, Qazvin, Iran.
amir@jadidi.info
[2]Database Research Group, University of Tehran, Tehran, Iran.
m.mohtarami@yahoo.com, h.amiri@ece.ut.ac.ir

In this research we applied Local Cluster Analysis (LCA) in tandem with Part-of-Speech tagging to monolingual task. We study different Persian POS tags and select a set of designated tags to reduce the size of our index and store the rich content of the documents. In addition, we applied LCA on the retrieved documents to detect the relevant and irrelevant documents to the user query. The clustering method is an important part in our approach. So we address the problem of building effective and meaningful clustering and evaluate different well-known and state of the art clustering methods for better efficiency and effectiveness in the proposed approach. LCA make valuable improvement against initial retrieval on Hamshahri corpus. Regarding CLEF train set, we get 26% improvement over MAP measure that compatible with same work on Hamshahri corpus but we have some technical problems with test set and results are weak.

Fusion of Retrieval Models at CLEF 2008 Ad-Hoc Persian Track

Zahra Aghazade*, Nazanin Dehghani*, Leili Farzinvash*, Razieh Rahimi*, Abolfazel AleAhmad*,
Hadi Amiri and Farhad Oroumchian**
* Department of ECE, University of Tehran
{z.aghazadeh, n.dehghany, l.farzinvash, r.rahimi}@ece.ut.ac.ir
** University of Wollongong in Dubai
FarhadO@uow.edu.au

Metasearch engines submit the user query to several underlying search engines and then merge their retrieved results to generate a single list that is more effective to the users' information needs. In this study, we try to use the idea behind metasearch engines in order to improve the results of Persian information retrieval. We consider each retrieval model as a decision maker and then fuse their decisions with an OWA operator in order to increase the effectiveness. We use an extension of Ordered Weighted Average (OWA) operator called IOWA and a weighting schema, NOWA for merging the results. Our experimental results show that merging by OWA operators produces better precision.

This was our first participation in CLEF as such it was not without mistake. We submitted 11 runs but instead of top 1000 retrieved documents, we reported on top 100 documents. This mistake has reduced the overall performance of the systems. We used nine different models from Terrier toolkit and then combined their results with NOWA and IOWA methods. Neither of the OWA methods showed any improvement over the original nine methods. After CLEF we did a further study and combined three systems with different token types, namely 4-grams, unstemmed single words and stemmed single words. On training set, IOWA and NOWA methods have shown 10% and 8.7% improvements over the original systems average precisions. However, on the test set these improvements were reduced to 5.6%.

Based on the preliminary results obtained, we believe the right way of using fusion on Persian language is combining systems with different token types. In future, we are going to study the effect of different token types and retrieval engines on fusion.

Robust-WSD

SINAI at Robust WSD Task @ CLEF 2008: When WSD is a Good Idea for Information Retrieval tasks?

Fernando Martínez-Santiago, José M. Perea-Ortega and Miguel A. García-Cumbreras
SINAI Research Group. Computer Science Department. University of Jaén
Campus Las lagunillas, Ed. A3, E-23071, Jaén, Spain
{dofer,jmperea,magc}@ujaen.es

SINAI has participated in the first edition of Robust WSD task with the aim of investigating the performance of disambiguation tools applied to Information Retrieval (IR). The main interest of our experimentation is the characterization of queries where WSD is a useful tool. That is, which issues must be fulfilled by a query in order to apply a state-of-art WSD tool? In the experiments carried out, we have used the two disambiguated collections provided by the NUS and UBC teams and the default collection for Robust WSD task without WSD data. After the interpretation of our experiments, we think that only queries with terms very polysemous and very high IDF value are improved by using WSD. We find that there are situations where WSD must be used, but these scenarios are very specific.

Uniba-Sense at Clef 2008: Semantic N-Levels Search Engine

Pierpaolo Basile, Annalina Caputo and Giovanni Semeraro
Department of Computer Science - Univerisity of Bari (Italy)
{basilepp,acaputo,semeraro}@di.uniba.it

We present evaluation experiments conducted at the University of Bari for the Ad-Hoc Robust WSD task of the Cross-Language Evaluation Forum (CLEF) 2008. The evaluation was performed using SENSE (SEmantic N-levels Search Engine), a system for text retrieval based on N-levels model. The system tries to overcome the limitations of the ranked keyword approach, by introducing semantic levels, which integrate (and not simply replace) the lexical level represented by keywords. In our interpretation, the semantic level provides information about word meanings, as described in a reference dictionary. We show how SENSE is able to manage documents indexed at two separate levels, keywords and word meanings, as well as to combine keyword search with semantic information provided by the other indexing level. We provide a detailed description of the SENSE model by defining a local scoring function, a local similarity function for synsets and a global ranking function in order to merge rankings produced by different levels, in an attempt of improving the retrieval performance. Experiments have been carried on the Ad-hoc Robust WSD CLEF 2008 dataset in order to evaluate the effectiveness of our approach. Results obtained by combining keywords and word meanings extracted from the WordNet lexical database, show the promise of the idea. In particular they confirm our hypothesis: The combination of keyword and meaning levels is more effective than the single keyword level. We obtain an improvement of 35% in precision using the N-levels model with respect to the keyword level alone. Moreover, the Precision-Recall curve shows that the N-levels model outperforms keyword level at all values of recall. As regards the CLEF competition, our system has a low precision with respect to the other participants. This is due to the standard relevance function implemented in Lucene, the API that we use to implement our model, and this result was expected. Lucene performance decreases when the number of terms in a query grows. This problem was discussed by other participants to the previous edition of TREC conference. The goal of our evaluation was to prove the effectiveness of the N-levels model and all the experiments confirmed our hypothesis. As future research, we plan to improve the performance of the system. This goal can be achieved by adopting two different strategies: The former involves the change of the Lucene relevance function; the latter exploits the possibility to replace vector space model with a more effective IR model.

IXA at CLEF 2008 Robust-WSD Task: Using Word Sense Disambiguation for (Cross Lingual) Information Retrieval

Arantxa Otegi, Eneko Agirre and German Rigau
IXA NLP Group - University of the Basque Country
Donostia, Basque Country
aotegui004@ikasle.ehu.es

This paper describes the participation of the IXA NLP group at the CLEF 2008 Robust-WSD Task. This is our first time at CLEF, and we participated at both the monolingual (English) and the bilingual (Spanish to English) subtasks. We tried several query and document expansion and translation strategies, with and without the use of the word sense disambiguation results provided by the organizers. All expansions and translations were done using the English and Spanish wordnets as provided by the organizers and no other resource was used. We used Indri as the search engine, which we tuned in the training part. Our main goal was to improve (Cross Lingual) Information Retrieval results using WSD information, and we attained improvements in both mono and bilingual subtasks, although the improvement was only significant for the bilingual subtask. As a secondary goal, our best systems ranked 4th overall and 3rd overall in the monolingual and bilingual subtasks, respectively.

UFRGS@CLEF2008: Indexing Multiword Expressions for Information Retrieval

Otavio Costa Acosta, André Pinto Geraldo, Viviane Moreira Orengo and Aline Villavicencio
Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
[ocacosta, apgeraldo, vmorengo, avillavicencio]@inf.ufrgs.br

For UFRGS's participation on CLEF's Robust task, our aim was to assess the benefits of identifying and indexing multiword expressions (MWEs) for Information Retrieval (IR). MWEs are sequences of words that act as a single unit for the purpose of linguistic analysis. The meaning of the expression is different from the meaning of its composing terms analysed individually. The correct identification and treatment of MWEs is important for IR since in an ideal IR system, the entries in the index should represent the concepts present in the documents. Indexing a MWE as separate terms will mean loss in semantics.

The approach used to identify MWEs was totally statistical, based association measures such as Mutual Information and Chi-square. These measures were applied over pairs of adjacent words, known as bigrams. We focused only on bigrams which have nouns (NN). Each time a MWE candidate was found in a document, we added the MWE candidate, joined by an underscore, to the document.

We worked on the English news collections composed by LA Times 94 and Glasgow Herald 95. Two versions of the collection were available: a "plain" version, and a version with word-sense disambiguation (WSD) data. Using the WSD documents (UBC version), we created a document collection composed by the lemmas in the texts. This collection was used as the basis for all our WSD runs. The IR system we used was Zettair. We have also used the Porter Stemmer.

Contradicting our results on the training topics, the results on the test topics did not show any significant improvements when MWEs were indexed. However, for some queries, the identification of MWEs was very important.

We also tested the opposite approach, i.e. removing all compounds from the texts. Since our collections were composed by lemmas, some terms were joined by an underscore, e.g. "to_have". These results also do not show any statistical differences.

In addition to the monolingual experiments, we also submitted four bilingual runs using Spanish topics to query English documents. The method used to map concepts between languages employed algorithms for mining association rules. Our bilingual experiments achieved 84% of their monolingual counterparts.

IRn in the CLEF Robust WSD Task 2008

Sergio Navarro, Fernando Llopis and Rafael Muñoz
Natural Language Processing and Information Systems Group. University of Alicante, Spain. Spain
{snavarro,llopis,rafael}@dlsi.ua.es

This paper describes our participation in the Robust WSD Task within the CLEF 2008. The aim of this pilot task is exploring methods which can take profit of WSD information in order to improve the IR systems. In our approach we have used a passage based system jointly with a WordNet based expansion method for the collection documents and the queries using the two WSD systems runs provided by the organization. Furthermore we have experimented with two well known relevance feedback methods - LCA and PRF -, in order to figure out which is more suitable to take profit of the WSD query expansion based on Wordnet. Our best run has obtained a 4th place in the competition with a value of 0.4008 MAP. We conclude that LCA fits better than PRF to this task. And that our WSD expansion is useful for some query subsets. In future works we will study the features of the query subsets for which the performance of our system decreases.

UCM-Y!R at CLEF 2008 Robust and WSD Tasks

José R. Pérez-Agüera[1] and Hugo Zaragoza[2]
[1]Universidad Complutense de Madrid (UCM)
[2]Yahoo! Research Y!R
jose.aguera@fdi.ucm.es, hugoz@yahoo-inc.com

Exploiting semantic information for information retrieval is known to be very hard. One of the problems, in our opinion, is the term independence hypothesis. A second problem is that of “query-dependant semantics”: two terms semantically related in a query may not be so in the next. We try to address these two problems. We propose to make explicit some of the term dependence information using a form of structured query (which we call query clauses), and to use a ranking function capable of taking the structure information into account. We combine the use of query expansion techniques and semantic disambiguation to construct the structured queries that are both semantically rich and focused on the query.

We explore the use of state of the art query expansion techniques combined with a new family of ranking functions which can take into account some semantic structure in the query. This structure is extracted from Wordnet similarity measures. Our approach produces improvements over the baseline and over query expansion methods for a number of performance measures including GMAP.

Simply adding terms to a query may not be the best way to enrich them. We believe that adding related terms worsens the term independence hypothesis. We have explored an alternative family of ranking functions that addresses this issue. These ranking functions and their motivation were described in more detail in (PerezAgueraZA08). Here we will give only a brief description.

Related terms are grouped in sets called clauses, and queries are defined as sets of clauses. Terms within the clauses and clauses themselves may be weighted. Each clause is considered as a pseudo term with each own tf and idf.

Our hypothesis is that semantically related terms should be grouped in clauses. The CLEF corpus is ideal to test this hypothesis since all the terms in it have be annotated with their corresponding synset in Wordnet

We can see that the proposed method improves results over the baseline and over query expansion, for all relevance measures including GMAP. This is very encouraging because it is one of the few results to our knowledge that show that semantic disambiguation can be used to improve retrieval in an open domain.

In our opinion a bottleneck to further improve performance is the difficulty of creating good query clauses. Wordnet Similarity methods tend to produce noisy clauses, often putting in correspondence terms that are not related in the context of the query.

UNIGE Experiments on Robust Word Sense Disambiguation

Jacques Guyot, Gilles Falquet, Saïd Radhouani and Karim Benzineb
Centre universitaire d'informatique, University of Geneva
Route de Drize 7, 1227 Carouge
jacques.guyot, gilles.falquet, said.radhouani@unige.ch; karim@alpineblue.eu

The aim of our experiments was to compare the results of two different retrieval techniques: the first one was based on the words found in documents and query texts; the second one was based on the senses (concepts) obtained by disambiguating the words in documents and queries. The underlying goal was to come up with more precise knowledge about the possible improvements brought by word sense disambiguation (WSD) in the information retrieval process. The proposed task structure was interesting in that it drew up a clear separation between the actors (humans or computers): those who provide the corpus, those who disambiguate it, and those who query it. Thus it was possible to test the universality and the interoperability of the methods and algorithms involved.

Intuitively, Word Sense Disambiguation should improve the quality of information retrieval systems. However, as already observed in previous experiments, this is only true in some specific situations, for instance when the disambiguation process is almost perfect, or in limited domains. The observations presented in our paper seem to support this statement. We propose two types of explanations:

1. When a query is large enough (more than one or two words), the probability that a document containing these words uses them with a meaning different from the intended one is very low. For instance, it is unlikely that a document containing mouse, cheese and cat is in fact about a computer mouse. This probably makes WSD useless in many situations. Such a request is similar in nature to the narrative-based tests. On the other hand, the WSD approach could make more sense when requests include only one or two words (which is the most frequent case in standard searches).
2. WSD is a very partial semantic analysis which is insufficient to really understand the queries. For instance, consider the query "Computer Viruses" whose narrative is "Relevant documents should mention the name of the computer virus, and possibly the damage it does". To find relevant documents, a system must recognize phrases which contain virus names ("the XX virus", "the virus named XX", "the virus known as XX", etc.). It should also recognize phrases describing damages ("XX erases the hard disk", "XX causes system crashes" but not "XX propagates through mail messages"). These tasks are very difficult to perform and they are far beyond the scope of WSD. Moreover, they require specific domain knowledge.

Ad Hoc Mixed: TEL, Persian & Robust

German, French, English and Persian Retrieval Experiments at CLEF 2008

Stephen Tomlinson
Open Text Corporation
stomlins@opentext.com

We describe evaluation experiments conducted by submitting retrieval runs for the monolingual German, French, English and Persian (Farsi) information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2008. In the ad hoc retrieval tasks, the system was given 50 natural language queries, and the goal was to find all of the relevant records or documents (with high precision) in a particular document set. We conducted diagnostic experiments with different techniques for matching word variations, comparing the performance on the robust Generalized Success@10 measure and the non-robust mean average precision measure. The measures generally agreed on the mean benefits of morphological techniques such as decompounding and stemming, but generally disagreed on the blind feedback technique, though not all of the mean differences were statistically significant. Also, 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.

The results suggest that, on average, the percentage of relevant items assessed was less than 55% for each of German, French and English and less than 25% for Persian.

JHU Ad Hoc Experiments at CLEF 2008

Paul McNamee
Johns Hopkins University
Human Language Technology Center of Excellence
paul.mcnamee@jhuapl.edu

For CLEF 2008 JHU conducted monolingual and bilingual experiments in the ad hoc TEL and Persian tasks.

The TEL task involved focused on searching electronic card catalog records in English, French, and German using data from the British Library, the Bibliotheque Nationale de France, and the Osterreichische Nationalbibliothek (Austrian National Library). The approach we adopted for TEL was to strip out non-content sections of records and to treat the task as ordinary full-text search using character n-grams and stemmed words.

For the Persian task, which is based on the Hamshahri corpus, several different forms of textual normalization were compared. Using the provided training topics we compared character n-grams, n-gram stems, ordinary words, words automatically segmented into morphemes, and a novel form of n-gram indexing based on n-grams with character skips. On the training topics we found that character 5-grams and skipgrams performed the best and this was borne out in our official submissions.

We also did some post hoc experiments using previous CLEF ad hoc tests sets in 13 languages.

In all three tasks we explored alternative methods of tokenizing documents including plain words, stemmed words, automatically induced segments, a single selected n-grams for each words, and all n-grams from words (i.e., traditional character n-grams). Character n-grams demonstrated consistent gains over ordinary words in each of these three diverse sets of experiments. Using mean average precision, relative gains of of 50-200% on the TEL task, 5% on the Persian task, and 18% averaged over 13 languages from past CLEF evaluations, were observed.

UniNE at CLEF 2008: TEL, Persian and Robust IR

Ljiljana Dolamic, Claire Fautsch and Jacques Savoy
Computer Science Department
University of Neuchatel, Switzerland
{Ljiljana.Dolamic, Claire.Fautsch, Jacques.Savoy}@unine.ch

In participating in the Ad Hoc evaluation campaign, our first objective is to analyze the retrieval effectiveness when using TEL (The European Library) corpora composed of very short descriptions (library catalogue records composed, in mean, from 12 (English) to 22 (German) indexing terms) and to evaluate the retrieval effectiveness of several IR models. The results of our various experiments demonstrate that the I(ne)B2 or PB2 models derived from the DFR paradigm or the LM model (for the German corpus) seem to provide the best overall retrieval performances. The Okapi model usually results in retrieval performances inferior to those obtained with the DFR or LM approaches.

The pseudo-relevance feedback (Rocchio) tends to hurt the MAP. The fact that the retrieved items are very short may explain this result. Therefore we do not recommend using blind query expansion with the TEL corpora. A data fusion strategy may enhance slightly the retrieval performance for the French or German corpus but hurt the retrieval performance with the English corpus. Therefore such a search strategy is also questionable in this context.

As a second objective we suggest and evaluate a stopword list and a light stemming strategy for the Persian language, a language having a relatively simple morphology. For this language, the I(ne)C2 tend to produce the best MAP. Moreover, our light stemmer tends to produce better MAP than does the 4-gram indexing scheme (relative difference around 5.5%). On the other hand, the performance difference with an approach ignoring a stemming stage is rather small (This result tends to indicate that the suffixes are already separated from their corresponding stem by a space). Using Rocchio's pseudo-relevance feedback, we may clearly improve the MAP. A data fusion strategy may also enhance the retrieval performance for the Persian corpus.

Finally, we participated in the robust track in an attempt to understand the difficulty involved in retrieving pertinent documents. Moreover, we made use of word sense disambiguation (WSD) information (lemma, POS tags, SYNSETS extracted from WordNet) to reduce problems related to polysemy when matching topic and document representation. Using blind query expansion and data fusion approaches (combining three IR models), we are able to improve the MAP from 0.4086 (Okapi) to 0.4515 (combined approaches), a relative improvement of 10.5%. However, the difference in MAP between runs with and without WSD information are rather small.

Mono- and Cross-Language Scientific Data Retrieval (Domain-Specific)

The Domain-Specific Track at CLEF 2008

Vivien Petras and Stefan Baerisch

GESIS Social Science Information Centre, Lennéstr. 30, 53113 Bonn, Germany

{vivien.petras | stefan.baerisch@gesis.org}

The domain-specific track evaluates retrieval systems for structured social science bibliographic collections in English, German and Russian.

The English and German GIRT databases contains from the German Social Science Information Centre's SOLIS (Social Science Literature) and SOFIS (Social Science Research Projects) databases from 1990-2000. CSA's Sociological Abstracts is provided as an additional English-language social science collection. The INION ISISS corpus covering social sciences and economics is used as Russian test collection. Documents contain textual elements (title, abstracts) as well as subject keywords from controlled vocabularies, which can be used in query expansion and bilingual translation.

In addition to the four test collections, various controlled vocabularies (thesauri) and mappings between vocabularies are available to participants. This year, new Russian language resources were provided, among them Russian-English and Russian-German terminology lists as well as a mapping table between the Russian and German controlled vocabularies.

25 topics were prepared in German and then translated into English and Russian. The retrieval tasks were monolingual, bilingual or multilingual retrieval against the German, English and Russian test collections.

Both feedback from the assessors as well as the precision numbers show that this year's topics were somewhat more difficult or more discriminating. The average number of relevant documents decreased for all three languages with Russian seeing the largest drop. As in previous years, the German and English averages are similar.

This year's track saw the use of a broad range of retrieval models, language processing, translation, and query expansion approaches. Statistical language models, probabilistic and vector-space models were employed with translation approaches that leverage thesaurus mappings as well as machine translation systems or web-based translation services. Two participants employed concept models based on semantic relatedness both for translation and query expansion.

Darmstadt (Müller & Gurevych, 2008) applied semantic models that utilize both Wikipedia and Wiktionary as sources for terms to form concepts that facilitate the use of semantic relatedness in the retrieval process. The Amsterdam (Meij & de Rijke, 2008) group used a language model approach to map between query terms, controlled vocabulary concepts and document terms. Berkeley (Larson, 2008) implemented a probabilistic logistic regression model with the Cheshire II system that was also employed for the Adhoc and GeoCLEF tracks. Chemnitz (Kürsten, Wilhelm & Eibl, 2008) used combinations of the Porter and the Krovetz stemmers for English and the Snowball stemmer and an N-Gram based decompounding approach for German and a stemmer developed by Unine for Russian, which worked well in combination with their Xtrieval framework IR system. The UniNE group (Fautsch, Dolamic & Savoy, 2008) tested four different blind feedback approaches. The classic Rocchio blind feedback method is compared to two variants of an approach that extends a query with terms selected based on their pseudo document frequency, which are considered for inclusion in the query if theyare within 10 words of the search term in the document. Finally, Google and Wikipedia were used for query expansion where the terms included in text snippets were used for query expansion. Geneva used their EasyIR system and the bilingual thesaurus for query expansion.

Pending availability of resources and demand, different tasks and options might be offered in 2009, e.g. additional corpus data (i.e. full text documents), a full topic run with 125 topics from the years 2003-2008, or other changes in the tasks.

UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches

Claire Fautsch, Ljiljana Dolamic and Jacques Savoy

Computer Science Department
University of Neuchatel, Switzerland
{Claire.Fautsch, Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

Our first objective in participating in this domain-specific evaluation campaign is to propose and evaluate various indexing and search strategies for the German, English and Russian languages, in an effort to obtain better retrieval effectiveness than that of the language-independent approach (n-gram). To do so we evaluate the GIRT-4 test-collection using the Okapi, various IR models derived from the Divergence from Randomness (DFR) paradigm, the statistical language model (LM) together with the classical tf.idf vector-processing scheme. For the German and Russian languages we applied our light stemming approach and stopword list. All our runs were fully automatic.

The resulting MAP show that the DFR I(n)B2 model usually provided in the best retrieval effectiveness for the German or English collections. The performance differences between Okapi and the various DFR models were usually rather small. For the Russian corpus, we found that when using word-based indexing, the DFR I(ne)B2 or the LM models tend to perform the best. With the 4-gram indexing approach, the LM model always presents the best performing schemes. Moreover, for the Russian language, this 4-gram indexing scheme tends to perform better than the word-based indexing strategy. Finally, the short query formulation (T) tends to produce a better retrieval performance than medium (TD) topic formulation.

In our analysis of Rocchio's blind query expansion approach, we find that this type of automatic query expansion can sometimes hurt the MAP or in other cases enhance it. For example this search strategy results in less improvement for the English corpus than it does for the Russian collection. For the German collection however, this search strategy clearly decrease the MAP.

This year we suggest two new query expansion techniques. The first, denoted "idf-window", is based on co-occurrence of relatively rare terms in a close context. As a second approach, we add the first two text snippets found by Google to expand the query. Compared to the performance before query expansion (e.g., German corpus, Okapi produces a MAP of 0.4096), Rocchio's blind query expansion cannot improve this retrieval performance. On the other hand, the new variant “idf-window” presents a better retrieval performance (+4.9%, from 0.4069 to 0.4247). Using the first two text snippets returned by Google, we may also enhance slightly the MAP (from 0.4096 to 0.4196, or +2.4%).

Back to Basics - Again - for Domain Specific Retrieval

Ray R. Larson
University of California, Berkeley - School of Information
ray@sims.berkeley.edu

In this paper we will describe Berkeley's approach to the Domain Specific (DS) track for CLEF 2008. Last year (2007) we used Entry Vocabulary Indexes and Thesaurus expansion approaches for DS, but found in later testing that some simple text retrieval approaches had better results than these more complex query expansion approaches. This year we decided to revisit our basic text retrieval approaches and see how they would stack up against the various expansion approaches used by other groups. The results are now in and the answer is clear, they perform pretty badly compared to other groups' approaches.

All of the runs submitted were performed using the Cheshire II system. This year the Berkeley/Cheshire group submitted a total of twenty-four runs, including two for each subtask of the DS track. These include six Monolingual runs for English, German, and Russian, twelve Bilingual runs (four X2EN, four X2DE, and four X2RU), and six Multilingual runs (two EN, two DE, and two RU). The overall results include Cheshire runs in the top five participants for each task, but usually as the lowest of the five (and often fewer) groups.

First Participation of University and Hospitals of Geneva to Domain-Specific Track in CLEF 2008

Julien Gobeill and Patrick Ruch
University and Hospitals of Geneva, Switzerland
julien.gobeill@sim.hcuge.ch

We participate in 2008 to our first Domain-Specific Track, with the aim to establish a baseline for our Information Retrieval engine in an unknown domain for us. We are specialized in Natural Language Processing in the biomedical domain, and we participate to the medical Image track and to TREC Genomics for four years with textual strategies, as queries expansions with controlled vocabularies, pattern recognition and vectorial space models. The technical component of our cross-language search engine is a generic toolkit, EasyIR, with which we can perform Text Categorization and Information Retrieval. The strategy applied for the 2008 Domain-Specific track is as simple as possible, as we want only to establish a baseline for EasyIR in a new track. For the English monolingual task, we choose to work with the title, the descriptive text and some types of classification terms to index documents. For the German queries to English collection bilingual task, we choose to perform a simple retrieval on the German collection in one hand, and to collect the descriptors of the retrieved documents in order to make cross-lingual query expansion in the other hand. Unfortunately, our results cannot be seen as fair, as we achieve MAP of 0.171 for the monolingual task and MAP of 0.132 for the bilingual task. Nevertheless, comparing to several baseline runs of other participants for DS CLEF 2007, our baseline run achieves equal performances. Possibilities to improve for the next DS CLEF are best tuning of our system with the benchmark, and an efficient use of the controlled vocabularies.

The University of Amsterdam at the CLEF 2008 Domain Specific Track: Parsimonious Relevance and Concept Models

Edgar Meij and Maarten de Rijke
ISLA, University of Amsterdam
{emeij, mdr}@science.uva.nl

We describe our participation in the CLEF 2008 Domain Specific track. The research questions we address are threefold: (i) what are the effects of estimating and applying relevance models to the domain specific collection used at CLEF 2008, (ii)what are the results of parsimonizing these relevance models, and (iii) what are the results of applying concept models for blind relevance feedback? Parsimonization is a technique by which the term probabilities in a language model may be re-estimated based on a comparison with a reference model, making the resulting model more sparse and to the point. Concept models are term distributions over vocabulary terms, based on the language associated with concepts in a thesaurus or ontology and are estimated using the documents which are annotated with concepts. Concept models may be used for blind relevance feedback, by first translating a query to concepts and then back to query terms. We find that applying relevance models helps significantly for the current test collection, in terms of both mean average precision and early precision. Moreover, parsimonizing the relevance models helps mean average precision on title-only queries and early precision on title+narrative queries. Our concept models are able to significantly outperform a baseline query-likelihood run, both in terms of mean average precision and early precision on both title-only and title+narrative queries.

The Xtrieval Framework at CLEF 2008: Domain-Specific Track

Jens Kürsten, Thomas Wilhelm and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media, 09107 Chemnitz, Germany
[jens.kuersten j thomas.wilhelm j maximilian.eibl] at cs.tu-chemnitz.de

This article describes our participation at the Domain-Specific track. We used the Xtrieval framework for the preparation and execution of the experiments. The translation of the topics for the cross-lingual experiments was realized with a plug-in to access the Google AJAX language API. This year, we submitted 20 experiments in total, whereof 5 were monolingual, 12 were bilingual and 3 multilingual runs. In all our experiments we applied a standard top-k pseudo-relevance feedback algorithm. Also, all of our submissions were merged experiments, where multiple stemming approaches for each language were combined to improve retrieval performance. The evaluation of the experiments showed that the combination of stemming methods works very well. Translating the topics for the bilingual experiments deteriorated the retrieval effectiveness only between 8 and 15 percent in comparison to our best monolingual experiments. A remarkably well performance was also achieved for all multilingual experiments. In our opinion the most interesting observation was that using the provided domain-specific thesauri did not improve retrieval performance. Finally, we would like to state that the strong performance of our cross-lingual experiments is most likely to be credited to the quality of the translation of topics.

Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval

Christof Müller and Iryna Gurevych
Ubiquitous Knowledge Processing Lab, Computer Science Department
Technische Universität Darmstadt,
Hochschulstr. 10, D-64289 Darmstadt, Germany
{mueller, gurevych}@tk.informatik.tu-darmstadt.de

The main objective of our experiments in the domain-specific track is utilizing semantic knowledge from collaborative knowledge bases such as Wikipedia and Wiktionary to improve the effectiveness of information retrieval (IR). While Wikipedia has already been used in IR, the application of Wiktionary in this task is new.

We evaluate two IR models, i.e. SR-Text and SR-Word, based on semantic relatedness by comparing their performance to a statistical model as implemented by Lucene. In both semantic models, the articles in Wikipedia and the word entries in Wiktionary are employed as textual representations of concepts. Each query or document term is then represented as a vector in the concept space according to its occurrence in the respective textual representation of the concept, i.e. in the article or word entry. The cosine of two concept vectors is used as a measure of semantic relatedness. The SR-Text model computes the similarity of a query and document using a centroid-based classifier. The SR-Word model combines individual similarities of each query and document term pair that are above a predefined threshold and then applies a set of heuristics.

In the monolingual task, we found that SR-Word outperforms SR-Text in most experiments. SR-Word outperforms Lucene only in one experiment. However, when Lucene is combined with the semantic models by using the CombSUM method, the mean average precision increases by 14% for German, 9% for English, and 16% for Russian.

In the bilingual task, we translate the English topics into the document language, i.e. German, by using machine translation. For SR-Text, we additionally explore a different method using the cross-language links between different language editions of Wikipedia. A cross-language link points from an article in one language to the same article in a different language. Using these links, we are able to map a concept vector whose concepts are represented by articles in the English Wikipedia into a concept vector whose concepts are represented by articles in the German Wikipedia. Thus, by transforming the concept vector of an English query using cross-language links, the similarity between the English query and German documents is computed by SR-Text without actually translating the query. This approach especially improves the retrieval performance in cases where the machine translation system incorrectly translates terms. When Lucene is combined with SR-Text, the mean average precision increases by 34%.

Interactive Cross-Language Retrieval (iCLEF)

Overview of iCLEF 2008:
Search Log Analysis for Multilingual Image Retrieval

Julio Gonzalo[1], Paul Clough[2] and Jussi Karlgren[3]
[1] UNED Spain
[2] University of Sheffield
[3] SICS Sweden

This paper summarises activities from iCLEF, the interactive track of CLEF (Cross-Language Evaluation Forum), an annual evaluation exercise for Multilingual Information Access systems. In iCLEF, cross-language search capabilities are studied from a user-inclusive perspective. A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language. Since 2006, iCLEF has moved away from news collections (a standard for text retrieval experiments) in order to explore user behaviour in scenarios where the cross-language search necessity arises more naturally for the average user. iCLEF has since based its experiments on Flickr, a large-scale, web-based image database. Flickr is based on a large social network of web users sharing over two billion images, with the potential for offering both challenging and realistic multilingual search tasks for interactive experiments.

In an attempt to encourage greater participation in user-orientated experiments, a new task was designed for 2008. The main novelty of the iCLEF 2008 task has been to focus experiments on a shared analysis of a large search log, generated by iCLEF participants from a single search interface provided by the iCLEF organizers. The focus is, therefore, on search log analysis rather than on system design. The idea is to study the behaviour of users in an (almost) naturalistic search scenario, having a much larger data set than in previous iCLEF campaigns. The search interface provided by iCLEF organizers is a basic cross-language retrieval system to the Flickr image database, presented as an online game: the user is given an image, and she must find it again without any a-priori knowledge of the language(s) in which the image is annotated. Game-like features are intended to engage casual users and therefore increase the chances of achieving a large, representative search log.

Six sites submitted results for this year's interactive track: Universidad Nacional de EducaciÃ³n a Distancia (UNED), the Swedish Institute of Computer Science (SICS), Manchester Metropolitan University (MMU), University of Padua (UNIPD), University of Westminster (UK), and the Indian Institute of Information Technology Hyderabad (IIIT-H).Studies ranged from exploring the effects of searcher background on results, studying how much attention searchers pay to language phenomena when searching images, how the effect of constraining the session might influence results, and examining logs to find evidence of user confidence in the search process. The results of the experiments will be used to inform more usage-oriented tasks for future cycles; the methodology has proven to be lightweight and should be helpful for future participants; the logs will be a sustainable and reusable resource for future studies.

FlickLing: a Multilingual Search Interface for Flickr

Víctor Peinado, Javier Artiles, Julio Gonzalo, Emma Barker and Fernando López-Ostenero
NLP & IR Group, ETSI Inform´atica, UNED
c/ Juan del Rosal, 16, E-28040 Madrid, Spain
victor@lsi.uned.es, javart@gmail.com, julio@lsi.uned.es, e.barker@dcs.shef.ac.uk, flopez@lsi.uned.es

This paper presents FlickLing, a multilingual search interface for Flickr designed and implemented by the UNED NLP & IR Group (nlp.uned.es) for the CLEF 2008 interactive task.

Flickling was designed to collect a large search log of multilingual image searches, which serves as the input data for the log analysis shared task at iCLEF 2008. FlickLing consists of two search modes (mono and multilingual) which allow to retrieve Flickr images annotated annotated in different languages. From a given query, FlickLing is able to automatically translate it into several languages (remembering the user's preferred term translations) and offer the user mechanisms to refine the query and improve the translations provided by the system. In addition, Flickling is offered to users as an online game with ranks for the best individual users and the best teams.

With the goal of collecting a large search log, Flickling works as an online competitive game, where users have to find as many images as possible to obtain the highest individual and team scores.

As target, Flickling users were given raw images (without annotations) and the goal was to find in the Flickr database as many images as possible in order to obtain the highest score for them and their teams. To do that, users can launch monolingual and multilingual searches, manipulate the automatic translations or refine their queries. When a user finds the target image, she obtains 25 points. At any time, users can quit and stop searching. When they do that, the system offers some hints to help finding the image. If users accept the hint, their score is penalised. Teams and users are ranked according to their score, precision (percentage of found images with respect to the images seen) and average time spent for each successful search session. The most challenging aspect of the task, besides the difficulty to describe the content of the picture and handling multiple languages, is that users don't know a priori which language(s) were used to annotate the image when it was uploaded into Flickr.

The paper describes the interface and the user logs generated, which has been used as a data source for the iCLEF 2008 log analysis task and contain over 5,000 complete search sessions made by over 200 users with a wide variety of language skills.

“Interactive” Undergraduate Students: UNIPD at iCLEF 2008

Giorgio Maria Di Nunzio
Department of Information Engineering - University of Padua
Via Gradenigo, 6/a - 35131 Padova - Italy
dinunzio@dei.unipd.it

This is the first year that the University of Padua (UNIPD) participates in the interactive CLEF (iCLEF) track. The iCLEF track is based on the Flickling search interface, a basic cross-language retrieval system for the Flickr image database, presented as an online game: the user is given an image, and he must find it again without any a priori knowledge of the language (one or more) in which the image is annotated.

In order to have a large number of participants, students of the Faculty of Humanities of the University of Padua, from the course of Linguistics and Modern Cultures and Languages for Cultural Mediation were asked to participate in the game. Participation was not mandatory; nevertheless, some incentive – extra points in the exam - was given in order to convince them to play. At the end, 60 students participated. The participation of these students was important for the aim of this study since these are users who use different languages every day.

From the analysis of the questionnaires and the system logs, interesting insights and results emerged and can be summarized with the following points:

- The hardest obstacle in finding the given image was probably the size of the set of images retrieved. In any case, image found or image skipped, a large number of users claimed that it was hard to find the image because there were too many images retrieved.

- Another hard point was the difficulty in describing the image. Finding suitable keywords is indeed a hard task. And images often have quite inappropriate tags or captions making it difficult for the students to find them. A solution to this problem could be adding the possibility to search according to visual features of the images. However, the answers in the questionnaires were not very positive about this tool.

- Users in general may find it difficult to describe the image because the language in which it is described is not known. As one could expect, this problem is less evident for the UNIPD language students. There is also the need for bilingual dictionaries with a better coverage, and for a system able to give good suggestions for translating the keywords.

- We also saw that there is no single strategy that outperforms the others. Using more monolingual searches than multilingual, a mix of the two, or preferring multilingual searches did not appear to influence the final score.

Evaluating the Impact of Personal Dictionaries for Cross-Language Information Retrieval of Socially Annotated Images

Diana Irina Tanase and Epaminondas Kapetanios
School of Computer Science,
University of Westminster,
London, UK

These working notes focus on the users' actions in order to assist translations and on the usage of personal dictionaries (a feature which enables saving user added words). The special interest for this feature comes from a need to investigate to what extent users get actively involved in the query translation and contribute to overcoming the limitations of automatic translations. It is also our hope that by understanding the relationship between user language skills and the usage of the personal dictionary feature in the iCLIR context, we will be able to get at least a partial answer to a bigger question regarding collaborative translations in today's participatory web space.

For analyzing the logs we have found the following answers to the research questions below:

1. Does the degree of confidence with a language affect usage and creation of personal dictionary entries, i.e., do those users with little knowledge of a language make use of the personal dictionary and to which extent?

The computed correlation between usage of the personal dictionary and language skill (-0.5946) indicates a decreasing linear dependency between the two.

2. Does the degree of confidence with a language affect quality of personal dictionary?

The log recorded a total of 460 new entries to the Personal Dictionaries. The most frequent are direct translations of the source query term, when there is no entry for it in the dictionaries. For the rest of the cases the users try to improve the provided translations list by adding synonyms, plural expressions, named entities, multiword expressions, or related concepts. Due to an average number of just 18 entries per user, it was hard to assess an overall trend for each of the five groups of users in terms of the quality of the personal dictionary.

3. Can it be inferred that the user's performance in the game results improved by using the personal dictionary and/or the assisted translation mechanism?

The language coefficient vs. distribution of translation related-actions showed a very weak correlation (-0.07266), while the same coefficient revealed a medium strength link between score and distribution of translation related-actions (0.3034); The correlation results also showed a very weak link between retrieval precision and distribution of translation related-actions (0.156).

4. Is the personal dictionary a useful interface facility?

The results for the overall questionnaire to the questions regarding the most useful interface facilities and the translation strategy showed that automatic and assisted translations were perceived as equally important features, while translations decisions were based on using known languages or other language resources outside the game.

UNED at iCLEF 2008: Analysis of a Large Log of Multilingual Image Searches in Flickr

Víctor Peinado, Julio Gonzalo, Javier Artiles and Fernando López-Ostenero
NLP & IR Group, ETSI Informática, UNED
c/ Juan del Rosal, 16, E-28040 Madrid, Spain
victor@lsi.uned.es, javart@gmail.com, {julio, flopez}@lsi.uned.es

In this paper, we summarise our analysis of the large log of multilingual image searches in Flickr provided to iCLEF 2008 participants.

In this search log, every session consists of a searcher (a registered user with a profile that includes her native language and her proficiency in English, Spanish, Italian, German, Dutch and French) and a target image (from the Flickr image database, annotated in one or more of that six languages). When the session starts, the user does not know in which language(s) the image is annotated. The interface provides facilities to perform queries simultaneously in up to six languages (via dictionary translation of query terms), to provide controlled relevance feedback (clicking on suggested terms and terms from the images found) and to refine the translations provided by the system (changing the selection of the system or adding new translations). The task is, therefore, a multilingual known-item retrieval task. If the user gives up, she can ask for hints; the first hint is the target language (which turns the task into bilingual or monolingual search, depending on the language profile of the user). The rest of the hints are keywords used to annotate the image, which is aimed at preventing users from being discouraged with difficult images.

The log consists of more than 5,000 search sessions by more than 200 users with a wide range of skills in the interface languages, coming from four continents. The size of this corpus permits studying the behaviour of users in a multilingual search scenario at a scale that had not been possible before.

The UNED team has focused on studying (a) correlations between the language skills of searchers in the target language and other session parameters, such as success (was the image found?), number of query refinements, etc.; (b) learning effects over time; (c) usage of specific cross-language search facilities and (d) users perceptions on the task (questionnaire analysis). This paper is a summary of our study.

We have identified 5101 complete search sessions (searcher/target image pairs) in the logs provided by the organisation. Our analysis shows that when users have active competence in the target language, their success rate is 12\% higher than if they do not know the language at all. If the user has passive competence of the language (i.e. can partially understand texts but cannot make queries), the success rate equals those with active competence, but at the expense of executing more interactions with the system.

The most remarkable learning effect is that users carry out fewer interactions when they are familiarised with the task and the system, keeping the success rate and the number of hints invariant. Finally, the usage of specific cross-language facilities (such as refining translations offered by the system) is low, but significantly higher than standard relevance feedback facilities, and is perceived as useful by searchers.

Finally, the perception of experience users about cross-language retrieval interactive facilities is very positive, in spite of the fact that they are not frequently used. This is an indication that advanced search features - in this case, manipulation of translations offered by the system - might not be used frequently, but when they are used they become critical for the success of the task. A consequence is that query translation assistance should be hidden in the default settings of a cross-language search interface, but should be possible to invoke it for certain advanced users or specific search situations.

SICS at iCLEF 2008: User Confidence and Satisfaction Inferred from iCLEF Logs

Jussi Karlgren
SICS
jussi@sics.se

SICS has participated in this year's iCLEF cycle mainly with an eye on future experimental settings to work on measurement of searcher trust and confidence in search results, in keeping with previous experimental studies performed at SICS to understand how to measure and assess trust and confidence. SICS has used the Flickling interface and the logs delivered by it to study how user actions can be interpreted as exponents of user confidence.

Variables under consideration for this purpose can be indirect, such as length of interaction, time spent on query formulation, and other measures which require non-trivial interpretation during the analysis phase. Other variables can be more direct, in that they more clearly indicate competence or confidence on the part of the user, such as observed edits and additions made to the user dictionaries by the user. Some such measures can be found in the logs, making a distinction between sequences of actions that ultimately provide a successfully identified target image and sequences which terminate by the user requesting another target.

For next year's cycle we will suggest more explicit measures of user confidence, as well as a task which better teases out differences between users' task termination decisions.

A Study of Users' Image Seeking Behaviour in Flickling

Evgenia Vassilakaki, Frances Johnson, R.J. Hartley and David Randall
Dept. Information & Communications
Manchester Metropolitan University
evgenia.vassilakaki@student.mmu.ac.uk, {f.johnson, r.j.hartley, d.randall}@mmu.ac.uk

This study aims to explore users' image seeking behaviour when searching for a known, non-annotated image in Flickling provided by iCLEF2008 track. The task assigned to users was to search for the three first images given after first login. Users did not know in advance in which of the six languages (English, German, Dutch, Spanish, French, Italian) the images were described, forcing them to search across languages. The main focus of our study was threefold: a) to identify the reasons that determined users' choice over a specific interface, b) to examine whether users were thinking about languages when searching for images and to what extent and c) to examine if used, how helpful the translations proved to be for finding the images.

This study used four different, both quantitative and qualitative methods (questionnaires, retrospective thinking aloud, observation and interviews) to meet its research questions. Results show that two out of ten users were using only the monolingual interface because they did not feel confident with languages and the rest were switching between interfaces for a variety of reasons in which languages played a small part. Only four out of ten users were actually thinking about languages when searching for the images, while the rest were more preoccupied with finding the images and completing the task successfully. As a consequence, only four users paid attention to translations and only judged the translations in languages known to them. Overall, the translations were not considered to be helpful due to their inconsistency in coverage and their tendency to lead to irrelevant results.

Mining the Behaviour of Users in a Multilingual Information Access Task

Srinivasarao Vundavalli
SIEL, LTRC, IIIT
Hyderabad, India
srinivasarao@research.iiit.ac.in

This paper summarizes the participation of IIIT-H in the CLEF 2008 interactive task. Our goal was to mine the logs and extract conclusions about the behavior of users when facing a strictly multilingual information access task. We are provided the search logs which are generated by an online game, known-item image retrieval from Flickr. In this paper we describe the following tasks. We looked for the differences in the search behavior according to the language skills. We clustered the users based on the score of the user, precision of the user and the number of hints he asked for. We then studied the behavior of the most successful user cluster, the least successful (unsuccessful) user cluster and the users in between the above two. Our results show that, most of the users start with monolingual interface and soon they realize cross-lingual is interface is more useful than mono-lingual interface, and the users are more comfortable to search in their mother language or the languages that they know.

Multiple Language Question Answering (QA@CLEF)

Overview of the Clef 2008 Multilingual Question Answering Track

Pamela Forner[1], Anselmo Peñas[2], Iñaki Alegria[3], Corina Forăscu[4], Nicolas Moreau[5], Petya Osenova, Prokopis Prokopidis[7], Paulo Rocha[8], Bogdan Sacaleanu[9], Richard Sutcliffe[10] and Erik Tjong Kim Sang [11]

[1] CELCT, Trento, Italy (forner@celct.it)
[2] Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain (anselmo@lsi.uned.es)
[3] University of Basque Country, Spain (i.alegria@ehu.es)
[4] Al. I. Cuza, University of Iasi, Romania Institute for Computer Science, Romania (corinfor@info.uaic.ro)
[5] ELDA/ELRA, Paris, France (moreau@elda.org)
[6] BTB, Bulgaria, (petya@bultreebank.org)
[7] ILSP Greece, Athena Research Center (prokopis@ilsp.gr)
[8] Linguateca, DEI UC, Portugal, (Paulo.Rocha@di.uminho.pt)
[9] DFKI, Germany, (bogdan@dfki.de)
[10] DLTG, University of Limerick, Ireland (richard.sutcliffe@ul.ie)
[11] University of Groningen (e.f.tjong.kim.sang@rug.nl)

This year the same evaluation setting as in 2007 campaign was proposed. Last year the task was changed considerably and this affected the general level of results and also the level of participation. Thus, the only differences from last year concerned the languages involved: as a compensation for the loss of Indonesian, two new languages (Basque and Greek) were introduced. Another important innovation was the possibility of returning more than one answer per question (up to three). Similarly to last year, every question was characterized by a topic, and Questions were grouped in clusters. The topics were not given to the systems, which had to infer them from the first question\answer pair. Questions in the same cluster could be linked through anaphoric coreferences. As last year, the systems were given the possibility to search for answers in Wikipedia (dumped at November 2006) as a document corpus, beside the usual newswire collections used in previous campaigns. This year participation increased slightly but the task proved to be still very difficult.

In addition to the main task, three additional exercises were offered, namely 1) the Answer Validation Exercise (AVE), 2) the Question Answering on Speech Transcriptions (QAST), which continued last year's successful pilot, and 3) Word Sense Disambiguation for Question Answering (QA-WSD), a pilot task which provided the questions and collections with already disambiguated Word Senses in order to study their contribution to the QA performance.

After reaching a high of 30 in 2006, the number of participants in the main track equalled that of last year i.e. 22. Nevertheless the number of submitted runs increased to 51 (31 for monolingual tasks and 20 for bilingual ones) compared to last year's 37 (23 in monolingual tasks and 14 in bilingual ones). As in previous campaigns, more participants chose the monolingual tasks.

As a general remark, it can be said that the innovations introduced in 2007 and continued in 2008 appeared to be a deterrent to some participants, who felt there was not enough time to update their systems to the new requirements. Moreover, the task proved to be still difficult even for veterans, probably due to the high level of difficulty of the question sets.

As far as evaluation is concerned, the traditional procedure was applied. Human judged assessed the exact answers as R (Right); W (Wrong); X (ineXact) and U (Unsupported). The main evaluation measure was Accuracy; MRR and Confidence Weighted Score, already used in the previous campaigns, were also calculated as auxiliary measures for systems which provided also a confidence score. As far as accuracy is concerned, scores were generally lower than usual. In fact, although best accuracy in the monolingual task increased with respect to last year, going up again to the values recorded in 2006, the systems - even those that have participated in all previous campaigns – did not achieve a brilliant overall performance. More in detail, best accuracy in the monolingual task scored 63.5% almost ten points up with respect to last year, meanwhile, the overall performance of the systems was quite low, as average accuracy was 23.63%, practically the same as last year. On the contrary, the performances in the cross-language tasks recorded a drastic drop: best accuracy reached only 19% compared to 41.75% in the previous year, which means more than 20 points lower, meanwhile average accuracy was more or less the same as in 2007 – 13.24% compared to 10.9%.

Overview of the Answer Validation Exercise 2008

Álvaro Rodrigo, Anselmo Peñas and Felisa Verdejo
Dpto. Lenguajes y Sistemas Informáticos, UNED
{alvarory, anselmo, felisa}@lsi.uned.es

The Answer Validation Exercise (AVE) at CLEF is aimed at developing systems able to decide whether the answer of a Question Answering (QA) system is correct or not. 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 (AV) is expected to be useful for improving QA systems performance. However, the evaluation methodology in AVE 2006 did not permit to quantify this improvement and thus, the exercise was modified in AVE 2007.

In AVE 2007 participant systems had to emulate QA systems selecting one answer per question from a set of candidate ones. These candidate answers were the ones given by QA systems participating at the QA main track at CLEF. This allowed us to study the use of AV systems as the answer selection method used by a QA system. Nevertheless, it was not acknowledged the ability of an AV system detecting if all the candidate answers to a question were incorrect. Systems with this ability could ask for new answers to the QA systems, opening the possibility of obtaining a correct answer to the question. Then, we have studied this behavior in AVE 2008 introducing new measures.

9 groups have participated with 24 runs in 5 different languages (German, English, Spanish, French and Romanian). Results show that AV systems could improve the performance of current QA systems when AV systems are used for selecting the final answer from a set of candidate ones. In fact, according to the results, except in the languages where the best QA system outperforms the others QA systems in more than a 50%, there was an AV system with better performance than QA systems. Besides, the use of the new measures that reward the detection of incorrect answers has given a more informative estimation of the potential of AV systems in QA performance.

The most used technique by the participants continues being lexical processing while the use of syntactic analysis has grown. Nevertheless, very few systems have performed semantic analysis. Besides, a high percent of participants have combined different features using Machine Learning. Finally, the best systems performed both lexical and syntactic analysis, taking into account Named Entities.

Overview of QAST 2008

Jordi Turmo[1], Pere Comas[1], Sophie Rosset[2], Lori Lamel[2], Nicolas Moreau[3] and Djamel Mostefa[3]
[1]TALP Research Centre (UPC). Barcelona. Spain
{turmo,pcomas}@lsi.upc.edu
[2]LIMSI. Paris. France
{rosset,lamel}@limsi.fr
[3]ELDA/ELRA. Paris. France
{moreau,mostefa}@elda.org

The objective of the QAST (QA in speech transcripts) track is to develop a framework in which QA systems can be evaluated when the answers have to be found in speech transcripts, these transcripts being either produced manually or automatically. The main objectives of the second edition of QAST to this evaluation are: motivating and driving the design of novel and robust QA architectures for speech transcripts, measuring the loss due to the inaccuracies in state-of-the-art ASR technology, measuring this loss at different ASR performance levels given by the ASR word error rate, comparing the performance of QA systems on different kinds of speech data, and motivating the development of monolingual QA systems for languages other than English.

A total of ten tasks were defined for this second edition of QAST covering five main task scenarios, three languages and different word error rates for automatic transcriptions (from 10.5% to 35.4%): lectures in English about speech and language processing, meetings in English about design of television remote controls, French broadcast news, and European Parliament debates in English and Spanish. The data for these tasks is derived from five different resources, covering spontaneous speech, semi-spontaneous speech and prepared speech: the CHIL corpus (lectures in English), the AMI corpus (meetings in English), the ESTER corpus (French broadcast news), the EPPS EN (European Parliament debates in English) and the EPPS ES (European Parliament debates in Spanish). Two types of questions were considered this year: factual questions and definitional ones. For each corpus, roughly 70% of the questions are factual, 20% are definitional, and 10% are NIL.

Each participant could submit up to 32 submissions (2 runs per task and transcription), and a total of 49 submissions from five groups of four different countries were evaluated. The results of the evaluated runs showed that for the tasks where the word error rate was low enough (around 10%) the loss in accuracy compared to manual transcriptions was under 5%, suggesting that QA in such documents is potentially feasible. However, even where ASR performance is reasonably good, there remain outstanding challenges in dealing with spoken language and the earlier mentioned differences from written language. The results indicate that if a QA system which performs well on manual transcriptions it also performs reasonably well on high quality automatic transcriptions. However, the performance on spoken language has not yet reached the level of those in the main QA track.

Monolingual and Bilingual QA

Experiments with Query Expansion in the RAPOSA (FOX) Question Answering System

Luís Sarmento, Jorge Teixeira and Eugénio Oliveira
Faculdade de Engenharia da Universidade do Porto
{las, teixeira.jorge, eco}@fe.up.pt

In this paper we present the results of applying a statistical query expansion method on the retrieval stage of a QA system for Portuguese (RAPOSA). Our approach involves expanding queries for event-related or action-related factoid questions using a verb thesaurus automatically generated using information extracted from a large web corpus.

Results on the monolingual Portuguese QA track show that our expansion approach improves QA recall when compared with applying expansion based on a simple form of stemming, while simultaneously requiring the analysis of only 30% as many text snippets. However, we were not able to outperform the recall obtained using an even simpler expansion method, which nevertheless achieves lower precision and requires analyzing many more text snippets.

We conclude by observing that a more thorough analysis of the usefulness of our approach on QA performance requires improving other stages of the QA pipeline, which currently impose significant limitations on the overall performance of the system.

AliQAn, Spanish QA System at CLEF-2008

S. Roger, K. Vila, A. Ferrández, M. Pardiño, J. M. Gómez, M. Puchol-Blasco and J. Peral
Natural Language Processing and Information Systems Group.
Department of Software and Computing Systems.
University of Alicante, Spain.
{sroger,kvila,antonio,maria,jmgomez,marcel,jperal}@dlsi.ua.es

This paper describes the participation of the system AliQAn, a monolingual open- domain Question Answering (QA) System developed in the Department of Language Processing and Information System at the University of Alicante, in the CLEF-2008 Spanish monolingual QA evaluation task. This year, the main contributions were: 1) Algorithm for resolving topic-related questions. The essence of this algorithm is to extend every question (qi) by adding some noun phrases and the noun of the answer of the first question of the same cluster which qi depends on. 2) Approach to decrease the number of inexact answers. This approach assigns certain weight (determined by using specific dictionaries) to the heads of each answer's noun phrase according to an expected answer type and it returns the head of the noun phrase with the greatest weight. We have obtained excellent results with a decrease of 20 inexact answers with regard to the year 2005. 3) Using Wikipedia with RI and QA systems. On one hand, our IR system has been adapted for making possible the use Wikipedia with very large document collections. On the other hand, several problems derived from the codification of the non-latin characters in Wikipedia have been resolved in order to use it together with our QA system.

Our system has treated all questions given in this track, except the list questions, and only one has been unsupported. Our paper only includes one run for the Spanish monolingual QA task and it has achieved an overall accuracy of 19.50%. Finally, we would like to point out that this is the first time we deal with Wikipedia and topic-related questions for our participation in the CLEF QA task.

Question Answering with Joost at CLEF 2008

Gosse Bouma, Jori Mur, Gertjan van Noord, Lonneke van der Plas and Joerg Tiedemann
Information Science, University of Groningen
g.bouma@rug.nl

We participated in the CLEF 2008 monolingual Dutch and multilingual English-to-Dutch question answering task.

Our QA-system, Joost, is an open-domain QA-system for Dutch, which makes heavy use of syntactic information in all of its components. The text collections used for CLEF (Dutch newspaper text and Wikipedia) are parsed using the Alpino parser, which performs part-of-speech and named entity tagging, and syntactic analysis using dependency relations. We use linguistic information for question analysis, for relation extraction, for building the IR-index, and for searching and ranking potential answer strings.

In 2008, we experimented with information extraction from Wikipedia infoboxes, with query expansion, and with an approach to MLQA based on Google Translate. The best monolingual run achieved an accuracy of 25.5%, and the best multilingual run achieved an accuracy of 13.5%.

IdSay: Question Answering for Portuguese

Gracinda Carvalho[1], David Martins de Matos and Vitor Rocio
[1] Universidade Aberta L2F/INESC-ID Lisboa - CITI – FCT/UNL, Rua da Escola Politécnica, 147, Lisboa, Portugal
[2] L2F/INESC-ID Lisboa, Instituto Superior Técnico/UTL, Rua Alves Redol 9, Lisboa, Portugal
[3] Universidade Aberta, CITI – FCT/UNL, Rua da Escola Politécnica, 147, Lisboa, Portugal
{gracindac, vjr}@univ-ab.pt, david.matos@inesc-id.pt

IdSay was submitted to the monolingual Portuguese task of the Question Answering track of the Cross-Language Evaluation Forum 2008 (QA@CLEF) for the first time. It is an open domain question answering system for Portuguese that was developed from scratch, with the objective of optimizing resources, so that response time could be short. Its current version can be considered a baseline version, using mainly techniques from the area of Information Retrieval. The only external information that it uses besides the text collections is lexical information for Portuguese. The index files for the text collection occupies 1.15 GB of disk space, and took about 4 hours to build. The load time is around 1 minute, and the time to process 200 questions is less than 1 minute.

At the QA@CLEF 2008 evaluation campaign it answered correctly to 65 of the 200 questions in the first answer, and to 85 answers, considering the three answers that could be returned per question, which correspond to an accuracy over the first question of 32.5% and overall accuracy of 42.5% and MRR over all questions of 37.083%.

Generally, the types of questions that are answered better by IdSay system are measure factoids (with an accuracy of 75.0%), count factoids (with an accuracy of 68.4%) and definitions (with an accuracy of 64.3%), but there is still work to be done in these areas, as well as in the treatment of time. List questions (that were not treated by the system) and location and people/organization factoids are the types of question with more room for evolution. The NIL accuracy (16.7%) for IdSay indicates a need of improvement in our mechanism to determine how well a passage supports the answers.

Cross Lingual Question Answering Using QRISTAL for CLEF 2008

Dominique Laurent, Patrick Séguéla and Sophie Nègre
Synapse Développement
Toulouse, France
{dlaurent, p.seguela, sophie.negre }@synapse-fr.com

QRISTAL is a question answering system making intensive use of natural language processing both for indexing documents and extracting answers. It ranked first in the EQueR evaluation campaign (Evalda, Technolangue) and in first rank in French for CLEF 2005, 2006 and 2007 [11], [12], [14]. This article describes the improvements of the system since last year. Then, it presents our benchmarked results for the CLEF 2008 campaign and a critical description of the system. Since Synapse Développement is participating to Quaero project, QRISTAL is most likely to be integrated in a mass market search engine in the forthcoming years.

Priberam's Question Answering System in QA@CLEF 2008

Carlos Amaral, Adán Cassan, Helena Figueira, André Martins, Afonso Mendes, Pedro Mendes, José Pina and Claudia Pinto
Priberam
Lisboa, Portugal
{cma, ach, hgf, atm, amm, prm, jfp, cp}@priberam.pt

This paper describes the major refinements implemented in Priberam's question answering (QA) system since our last CLEF participation, followed by the discussion of the results obtained in the Portuguese and Spanish monolingual runs at the main task of QA@CLEF 2008.

The performance of Priberam's system in last year's QA@CLEF displayed internal and external changes. Internally, the system underwent several modifications, both in the Portuguese and in the Spanish modules, the most relevant one being the introduction of syntactic question processing. Externally, the CLEF organisation introduced topic-related questions (questions clustered around a common topic that might present anaphoric links between them) and added Wikipedia as a target document collection to the already existent newspaper corpora. As a result, there was a slight increase of the overall accuracy in the Spanish run and a significant decrease of the overall accuracy in the Portuguese run. Nevertheless, Priberam's system achieved a more accurate question categorisation, hence decreasing the number of wrong candidate answers, due to the introduction of syntactic parsing during question processing.

The main goal of Priberam's participation in QA@CLEF 2008, following the results of last year's evaluation, was to stabilize the system in order to achieve its potential performance. To enhance the performance in terms of speed, we improved the indexing process, optimized the way indexes are stored and parallelised the algorithms of the retrieval process. These changes led to substantial improvements on the speed of the retrieval system and allowed querying for question categories at sentence level and for ontology domains of the expected answer in document retrieval without speed penalties. The fine-tuning of the syntactic analysis, by using the phrases' core nodes as objects, allowed the system to more precisely match the pivots of the question with their counterparts in the answer, taking into account their syntactic functions. Additionally, new mechanisms in the QAP syntax were added to inform the retrieval system on what and how to look for documents/sentences in the collections. The module for named entity extraction was improved, by widening the coverage of detection and by adding new semantic features. Unfortunately, it was not yet possible to do on the Spanish rules the work that was done on the Portuguese ones. This is very clear in the difference between the Portuguese and Spanish results and it is a good measure of the improvements made on the Portuguese rules. From the analysis of the results, we conclude that the retrieval stage and the question analysis stage are performing very well for questions like those posed in CLEF, that QAPs need to broaden their coverage and that the work done for Portuguese this year must be ported to the Spanish rules.

Multi-lingual Question Answering Using OpenEphyra

Menno van Zaanen
Tilburg University
mvzaanen@uvt.nl

In this article we describe our submission to the Dutch-English QA@CLEF task. We took the publicly available OpenEphyra question answering system, which is an open-source English question answering system. This was turned into a multi-lingual variant by translating questions from Dutch to English using Systran's online-translation system. The current approach has some known problems, for example, we do not distinguish between factoid, lists, and definition questions (all questions are treated as factoid questions), OpenEphyra does not provide support text for answers (text in the document surrounding the answer is used as support text), temporal restrictions and anaphora are not handled at all. The amount of modifications of OpenEphyra required to run the experiment were such that due to time constraints only one experiment could be submitted. The original idea behind this research was to investigate the impact of the quality of the question analysis. In particular, we are interested in the difference between the analysis on the question in the source language and the question in the target language.

Dublin City University at Multilingual Question Answering CLEF 2008

Sisay Fissaha Adafre[1] and Josef van Genabith[2]
[1] National Center for Language Technology, School of Computing, DCU, Dublin
[2] IBM CAS Dublin
sadafre,josef@computing.dcu.ie

We describe our first participation in Multilingual Question Answering at CLEF 2008 using German and English as our source and target languages, respectively. The aim of the current exercise is to apply in house developed Natural Language Processing (NLP) tools in the development of a Question Answering system, and to test to what extent off-the-shelf tools, i.e. UIMA (Unstructured Information Management Architecture), can help speed up the development process.

The system is largely based on Information Extraction methods, with various filtering and re-ranking steps to pin point the correct answers. Our question answering system consists of the following core components: Question Analysis, Passage Retrieval, Sentence Analysis and Answer Selection. The system uses both shallow and deep NLP techniques. We used the TreeTagger for POS tagging and Chunking, and a treebank-based Lexical Functional Grammar(LFG) parser (developed in Dublin City University) for dependency parsing. The system also uses WordNet and Wikipedia as lexical resources. The system is built using UIMA as underlying framework.

Overall the best performing system returned only 16 exact answers, and 25 correct answers counting em unsupported answers. The web re-ranking component contributed significantly. The result without the web re-ranking component is disappointing. This is attributed to a number of problems. The main problem was lack of proper testing due to time constraints. This was compounded by an error introduced by a last minute change. Another major problem is, of course, the scope of our system. The system relies primarily on online methods which focus on a restricted class of named entities. Since it is an evolving system, we believe that its coverage will improve by adding more semantic categories.

Our future plan is to extend the types of question that can be handled, and improve the methods for those already implemented. Furthermore, we also need to improve the re-ranking algorithms. We would like to bring in more of the deep NLP methods into the re-ranking algorithm. Specifically, we would like to extend our dependency triple based scoring method to include the full LFG-based parse output. Finally, computation of the overall score is based on a simple linear combination of the individual scores ignoring their relative weights. In the future, we will use an ML based approach for computing the overall score using the individual evidences as features.

University of Wolverhampton at CLEF 2008

Iustin Dornescu, Georgiana Puscasu and Constantin Orasan
Research Group in Computational Linguistics, University of Wolverhampton
{I.Dornescu2, georgie, C.Orasan}@wlv.ac.uk

This article presents the participation of University of Wolverhampton in the Romanian to English Question Answering task at CLEF-2008. The objective for this year was to develop a question answering (QA) framework in which different modules are plugged in dynamically so that different processing components can be included as needed. The main components of our system deal with the three standard stages used in question answering: question processing, paragraph retrieval and answer extraction, and the system's cross-linguality is ensured by a term translator. The question processor analyses Romanian questions and produces a detailed representation of each question including the terms it contains. English translations are then generated for all question terms by exploiting information included in the Romanian and English WordNets, as well as aligned Wikipedia pages. They form the query that Lucene uses to extract English paragraphs which constitute the input for an answer extractor largely based on the one distributed with the OpenEphyra framework. The results indicate a small improvement in comparison with last year's performance. The overall accuracy achieved by our system was 18%, and the highest accuracy per question type was 66.67% for definition questions.

Esfinge at CLEF 2008: Experimenting with Answer Retrieval Patterns: Can They Help?

Luís Costa
Linguateca, SINTEF ICT, NORWAY
luis.costa@sintef.no

Esfinge is a general domain Portuguese question answering system which has been participating at QA@CLEF since 2004. It uses the information available in the “official” document collections used in QA@CLEF (newspaper text and Wikipedia), but additionally it also uses information from the Web as an additional resource when searching for answers. Where it regards the use of external tools, Esfinge uses a syntactic analyzer, a morphological analyzer and a named entity recognizer. This year an alternative approach to retrieve answers was tested: whereas in previous years, search patterns were used to retrieve relevant documents, this year a new type of search patterns was also used to extract the answers themselves. Besides that we took advantage of the main novelty introduced this year by QA@CLEF organization which was that the systems could return up to three answers for each question, instead of the single answer allowed in previous editions. This enabled the investigation about how good were the second and third best answers returned by Esfinge (when the first answer is not correct). The experiments revealed that the answer retrieval patterns created for this participation improve the results, but only for definition questions. Regarding the study of the three answers returned by Esfinge, the conclusion was that when Esfinge answers correctly a question, it does so usually with its first answer.

AliQAn, Spanish QA System at Multilingual QA@CLEF-2008

R. Muñoz-Terol, M.Puchol-Blasco, M. Pardiño, J.M. Gómez, S.Roger, K. Vila, A. Ferrández, J. Peral and P. Martínez-Barco
Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
rafamt,marcel,maria,jmgomez,sroger,kvila,antonio,jperal,patricio@dlsi.ua.es

In QA@CLEF 2008, we participate in monolingual (Spanish) and multilingual (English - Spanish) tasks. Specifically, in the research work, we will tackle with the English - Spanish QA task. In this edition we will deal with two main problems: an heterogeneous document collection (news articles and Wikipedia) and a large number of topic-related questions, which make somewhat difficult our participation. We want to highlight in the translation module in our system two possible mechanisms: one based on logic forms, and the other, on machine translation techniques. In addition, it has also been used a system of anaphora resolution that it is described below and a QA System, AliQAn (also used this year in the monolingual task). The scores obtained in the application of the machine translation techniques are a bit better than the ones obtained in the application of the techniques based on logic forms. This can be due to the fact that the use of logic forms is a good method to perform the language-independent knowledge representation, but this method must be improved to perform the translation of sentences from one language to another.

Ihardetsi Question Answering System at QA@CLEF 2008

Olatz Ansa, Xabier Arregi, Arantxa Otegi and Ander Soraluce
IXA Group. University of the Basque Country
olatz.ansa@ehu.es

This paper describes IHARDETSI, a monolingual question answering system for Basque and presents the results of our first participation in the QA@CLEF. We participated in the task using Basque, English and Spanish as source languages and Basque as target language. The main goal of our participation was to evaluate our basic system in order to compare with any other systems dealing with Basque and with the state of the art of non-English question answering systems.

Ihardetsi relies on NLP tools, which perform a linguistic analysis both on the question and on the corpus. We use a Spanish-Basque and an English-Basque machine translation system for the Cross-lingual tasks to translate the questions into Basque. The monolingual system has three main modules: question analysis, passage retrieval and answer extraction.

The question analysis module analyses the question (by using a lemmatizer and an entity recognizer) in order to obtain the question type, the expected answer type, the question focus and the query terms. When necessary, the semantic file of the question focus is obtained from the BasqueWN and used to refine the expected answer type.

The passage retrieval module receives as input the query terms from the previous module. Besides, if the question is not the first one of a topic then this module receives the search terms of the first questions of its topic as well as its three returned answers. Then, a sequence of queries is created by removing terms based on their IDF (Inverse Document Frequency) value. After this, relevant passages from the document collection are retrieved.

Finally, the answer extraction module analyses the obtained passages and searches for any answers that satisfy the expected answer type. The possible answer set is ordered and the first three answers are returned.

We have submitted four runs, one for Basque-Basque task, another one for English-Basque task and the other two ones for Spanish-Basque task (one of them uses synonymy expansion and the other one not).

As expected, the best results were obtained in the run of the monolingual task. On the other hand, although the Spanish-Basque machine translation system was better than the English-Basque one, the results were very similar. The use of synonymy expansion technique did not improve the results.

DFKI-LT at QA@CLEF 2008

Bogdan Sacaleanu, Günter Neumann and Christian Spurk
LT-Lab, DFKI, Saarbrücken, Germany
{bogdan, neumann, cspurk}@dfki.de

This Working Note shortly presents QUANTICO, a cross-language open domain question answering system for German and English document collections. The main features of the system are: use of preemptive off-line document annotation with information like Named Entities, sentence boundaries and pronominal anaphora resolution; online extraction of abbreviation-extension pairs and appositional constructions for the answer extraction; use of online translation services for the cross-language scenarios and of English as interlingua for language combinations not supported directly; use of redundancy as an indicator of good answer candidates; selection of the best answers based on distance metrics defined over graph representations. Based on the question type two different strategies of answer extraction are triggered: for factoid questions answers are extracted from best IR-matched passages and selected by their redundancy and distance to the question keywords; for definition questions answers are considered to be either the first sentence of description paragraphs in Wikipedia documents or the most redundant normalized linguistic structures with explanatory role (i.e., appositions, abbreviation's extensions). The results of evaluating the system's performance by QA@CLEF 2008 were as follows: for the German-German run we achieved a best overall accuracy (ACC) of 37%; for the English-German run 14.5% (ACC); and for the German-English run 14% (ACC).

QA@L2F, Second Steps at QA@CLEF

Luísa Coheur, Ana Mendes, João Guimarães, Nuno J. Mamede and Ricardo Ribeiro
L2F/INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
qa-clef@l2f.inesc-id.pt

This paper describes the participation of QA@L2F, the question-answering system from L2F/INESC-ID, at the QA track of CLEF in 2008.

Making intensive use of a Natural Language Processing chain (which includes, among others, a morphological analyzer, a disambiguation module, a disambiguation module, a multi-word recognizer, a chunker and named entities recognizer), QA@L2F is based on a three module approach to answer questions: corpora pre-processing ,where the information sources are processed and potentially relevant information is extracted; question interpretation, where the question is converted into a frame; and answer extraction, where different strategies are used to retrieve the final answer to the input question.

QA@L2F system was created in 2007 and had its first participation at CLEF, with results we considered auspicious. Nevertheless, having in mind the objectives of correcting some detected failures, increasing the percentage of questions the system deals with and correctly answers, and also experiment new techniques using the same processing tools, the system suffered modifications during this year: the question interpretation step was improved to better profit from the results of the Natural Language Processing chain; an anaphora solver module was introduced, which allowed us to answer some questions containing backwards references; finally, some other small improvements were done on the system, especially in the answer extraction module.

QA@L2F had 20% of precision at the competition this year, which represents an increase in the number of correct answers returned by the system of 6%, as compared to the last year results. The system highest accuracy values are on definition questions, in which it achieved 60.714% of precision.

UAIC Participation at QA@CLEF2008

Adrian Iftene[1], Ionuț Pistol[1] and Diana Trandabăț[1, 2]
[1]UAIC: Faculty of Computer Science, "Alexandru Ioan Cuza" University, Romania
[2]Institute for Computer Science, Romanian Academy Iasi Branch
{adiftene, ipistol, dtrandabat}@info.uaic.ro

This year marked UAIC 's third consecutive participation at the QA@CLEF competition, with continually improving results. The most significant changes to our system with regards to last year is the partial transition to a real-time QA system, consequences being the simplification or elimination of principal time-consuming tasks such as linguistic pre-processing. A brief description of our system and an analysis of the errors introduced by each module are described in the paper.

In order to build a real-time QA, we eliminated most time-consuming pre-processing steps (part-of-speech and Named Entities identification) and we kept at minimum the number of tools involved in this part. This proved to not have a major impact on our results, as they are significantly better than last year's.

The second important improvement was regarding information retrieval part, where Lucene queries were built in a specific way for Definition questions, and the searches were done in files with the same title as the entity that must be defined. We indexed the corpora in two ways: at paragraph level and at document level, and we kept both types of returned snippets. If the search of the answer in paragraph snippets is without success, we try to identify the answer in documents snippets.

The last main improvement was done at the answer extraction part, where we tried to build very specific patterns in order to identify the final answer. For example, the MEASURE type was divided in three subtypes SURFACE, LENGTH, and MEASURE. In this way, we improved the quality of the extraction module by specialising the patterns used. Also, in order to extract for definitions questions, we use a specialised Romanian grammar.

The precision of our system was 31% , with 19% better than the accuracy obtained last year. The main improvement was done at questions of type definition where we got 17 correct answers, compared to none in 2007. A more detailed analysis of our results can be provided when the official "golden" answers are provided.

The significant improvements shown this year combined with the major reduction in the processing time required by our system show promise regarding our goal, which is to migrate towards real-time QA.

RACAI's QA System at the Romanian-Romanian Multiple Language Question Answering (QA@CLEF2008) Main Task

Radu Ion, Dan Ştefănescu, Alexandru Ceauşu and Dan Tufiş
Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania
radu@racai.ro, danstef@racai.ro, aceausu@racai.ro, tufis@racai.ro

The Research Institute for Artificial Intelligence of the Romanian Academy (RACAI) is at the 3rd participation in the CLEF series of Question Answering competitions. This year (as in the previous one) we have focused on automatically answering questions in Romanian by searching their answers in Romanian Wikipedia documents. Thus we have participated in the Romanian-Romanian Main Task of QA@CLEF2008.

Our present system is based on the one that we have developed for the previous CLEF competition (Tufiş et al., 2008c). The main differences reside in an improved query formulation module and a completely redesigned answer extraction module which uses the results of a snippet selection and ranking component which did not exist in the 2007 version of the system. Our current architecture consists of the following:

3. **question analysis** in which the topic/focus articulation, question type and answer type are identified;
4. **query formulation** in which the translation from natural language question to the search engine syntactically well-formed query takes place;
5. **information retrieval** in which, using a **search engine**, the top *K* documents matching the query from the previous step are returned;
6. **snippet selection and ranking** in which, from the results of the search engine, the top *M* text snippets of a given word count are selected and ordered as to the likeliness to contain the correct answer;
7. **answer extraction** in which, using the sorted list of *M* snippets, answers candidates are extracted as syntactically well-formed substrings and returned as a ordered list in the decreasing order of the likeliness to be the correct answer to the user's question.

Both the test set and the document collection were preprocessed with the TTL POS tagger, lemmatizer and chunker (Ion, 2007) and the question analysis employed the linkage analysis supplied by LexPar (Ion, 2007). In order to rank the snippets from the documents returned by the search engine (a C# port of Lucene Boolean searching engine) we developed a lexical chains procedure which uses the Romanian WordNet (Tufiş et al., 2008a).

We have developed an alternate test set called the normalized test set which is derived from the official test set by replacing the referential expressions/pronouns found in questions with their proper referents. We thus tested our system on both test sets. Table 1 contains the performance measures of our two official runs on the official test set. The last row of the table displays the results on the normalized test set.

Table 1: The answer extraction accuracy over the two test sets

Runs	MRR	Coverage
ICIA081RORO (official test set) Right (R)	0.0683	0.095
ICIA081RORO (official test set) ineXact (X)	0.0691	0.09
ICIA082RORO (official test set) Right (R)	0.1233	0.155
ICIA082RORO (official test set) ineXact (X)	0.0633	0.08
SSR (normalized test set) Right (R)	0.1815	0.365

REFERENCES

Ion, R., Word Sense Disambiguation Methods Applied to English and Romanian, PhD thesis, Romanian Academy, Bucharest, 2007.

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A., Ştefănescu, D., Romanian WordNet: Current State, New Applications and Prospects, In Attila Tanács et al. (eds.): "Proceedings of the Fourth Global WordNet Conference (GWC 2008)", Szeged, Hungary, pp 441 – 452, 2008a.

Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A., RACAI's Question Answering System at QA@CLEF2007, In Carol Peters et al. (eds.): "Evaluation of Multilingual and Multi-modal Information Retrieval 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers", Lecture Notes in Computer Science, Springer-Verlag, 2008c. (in press).

The LogAnswer Project at QA@CLEF 2008: Towards Logic-Based Question Answering

Ingo Glöckner[1] and Bjoern Pelzer[2]
[1]Intelligent Information and Communication Systems Group (IICS)
University of Hagen, 59084 Hagen, Germany
ingo.gloeckner@fernuni-hagen.de
[2] Department of Computer Science, Artificial Intelligence Research Group
University of Koblenz-Landau, Universitaetsstr. 1, 56070 Koblenz
bpelzer@uni-koblenz.de

LogAnswer is a logic-oriented question answering system jointly developed by the AI research group at the University of Koblenz-Landau and by the IICS at the University of Hagen. The system was designed to address two notorious problems of the logic-based approach: Achieving robustness and acceptable response times. The main innovation of LogAnswer is its use of logic for simultaneously extracting answer bindings and validating the corresponding answers. In this way the inefficiency of the classical answer extraction/answer validation pipeline is avoided. The prototype of the system, which can also be tested on the web, demonstrates response times suitable for real-time querying. Emphasis was also placed on developing techniques for making the logic-based approach more robust against gaps in the background knowledge and against errors of linguistic analysis. To this end, the optimized deductive subsystem is combined with shallow techniques by machine learning. The same background knowledge as in the MAVE validator of the IICS presented at CLEF 2007 was used: 10,000 lexical-semantic relations (e.g. describing nominalizations), 109 logical rules, and a list of synonyms covering more than 111,000 lexical constants which is also utilized for determining the shallow features. Two monolingual runs of LogAnswer for German were submitted to QA@CLEF 2008. The results of 29 correct answers in the best run (accuracy: 0.145) indicate that further development of the current prototype is necessary. An error analysis shows that the linguistic processing and also the coreference resolution generally performed quite well. The rudimentary implementation of answer extraction based on the answer substitution determined by the prover must be improved, though, since extracted answers for appositions and constructions involving a defining verb are not reliable yet.

The MIRACLE Team at the CLEF 2008 Multilingual Question Answering Track

Ángel Martínez-González[2], César de Pablo-Sánchez[1], Concepción Polo-Bayo[2], María Teresa Vicente-Díez[1], Paloma Martínez-Fernández[1] and José Luís Martínez-Fernández[1,2]
[1] Universidad Carlos III de Madrid
[2] DAEDALUS - Data, Decisions and Language, S.A
amartinez@daedalus.es, cdepablo@inf.uc3m.es, cpolo@daedalus.es, tvicente@inf.uc3m.es, pmf@inf.uc3m.es, jmartinez@daedalus.es

The MIRACLE team is a consortium formed by three universities from Madrid, (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) and DAEADALUS, a small and medium size enterprise (SME). The MIRACLE team participated in the monolingual Spanish and cross-language French to Spanish subtasks at QA@CLEF 2008. For the Spanish subtask, we used an almost completely rebuilt version of our system, designed with the aim of flexibly combining information sources and linguistic annotators for different languages. To allow easy development for new languages, most of the modules don't make any language dependent assumptions. This language dependent knowledge is encapsulated in a rule language developed within the MIRACLE team. By the time of submitting the runs, work on the new version was still ongoing, so we consider the results as a partial test of the possibilities of the new architecture. Subsystems for other languages were not yet available, so we tried a very simple approach for the French to Spanish subtask: questions were translated to Spanish with Babylon, and the output of this translation was fed into our system. The results were an accuracy of 16% for the monolingual Spanish task and 5% for the cross-language task.

The Senso Question Answering System at QA@CLEF 2008

José Saias and Paulo Quaresma
Universidade de Évora, Portugal
{jsaias,pq}@di.uevora.pt

The University of Évora participation in QA@CLEF2008 was focused on the Portuguese monolingual task and was based on the updated Senso Question Answering System.

This system uses a local knowledge base, providing semantic information for text search terms expansion. The solver module uses two components to collect plausible answers: the logic and the ad-hoc solvers. The logic solver starts by producing a First-Order Logic expression representing the question and a logic facts list representing the texts information and then it looks for answers within the facts list that unify and validate the question logic form. The ad-hoc solver is designed for cases where the answer can be directly detected in the text. Then all the results are merged for answer validation. Some answer values are refused if they are not in accordance with the question type. Besides filtering, each answer weight may suffer an adjustment to a more reliable value. The Web redundancy can be exploited as a method for answer validation in QA. The idea is to measure a statement popularity or acceptance with a Web search and take that into account for an answer accuracy validation.

The system answers were classified as Right for 93 questions, which corresponds to an overall accuracy score of 46.50% (4.5% more than obtained last year). The system returned only 21 NIL answers, significantly less when comparing to 2007. Despite this appears to be a better value, the accuracy for the NIL question type went down from 10.81% to 9.52%. In the Factoids category the system had an accuracy of 40.74%, quite similar to last year. The best relative accuracy result was again on the Definition question type with 85.71%. Being the second time this QA system is used, the results are in line with the expected. The document retrieval process update and the text search query generation process led to the identification of more candidate documents, decreasing the number of NIL answers. Some ad-hoc solver's rules need an adjustment. Before applying our system to other source languages, there is some work to do in order to make the system components independent from the language. One possibility for a future participation is the submission of multiple answers per question. That can be accomplished with this system by selecting the N most weighted answers.

University of Hagen at QA@CLEF 2008: Efficient Question Answering with Question Decomposition and Multiple Answer Streams

Sven Hartrumpf, Ingo Glöckner and Johannes Leveling
Intelligent Information and Communication Systems (IICS)}
University of Hagen (FernUniversität in Hagen). 58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de

The German question answering (QA) system IRSAW (formerly: InSicht) participated in QA@CLEF for the fifth time. IRSAW was introduced in 2007, by integrating the deep answer producer InSicht, several shallow answer producers, and a logical validator.

InSicht realizes a deep QA approach: it transforms documents to semantic representations using a parser, draws inferences on semantic representations with rules, and matches semantic representations derived from questions and documents. InSicht was improved for QA@CLEF 2008 mainly in the following areas. The coreference resolver was trained on question series instead of newspaper texts in order to be better applicable for follow-up questions in question series. Questions are decomposed by several methods on the level of semantic representations. On the shallow processing side, the number of answer producers was increased from 2 to 4, by adding FACT and SHASE.

The answer validator introduced in the previous year was replaced with the faster RAVE validator designed for logic-based answer validation under time constraints. Using RAVE for merging the results of the answer producers, monolingual German runs and bilingual runs with source language English and Spanish were produced by applying a machine translation web service. An error analysis showed the main problems for the precision-oriented deep answer producer InSicht and the potential offered by the recall-oriented shallow answer producers.

Answer Validation Execise (AVE)

The Answer Validation System ProdicosAV

C. Jacquin, L. Monceaux and E. Desmontils
Université de Nantes - Laboratoire LINA – France

In this paper, we present the ProdicosAV answer validation system which was developed by the TALN team from the LINA institute and which participated to the Answer Validation Exercice for French. This system is based on the Prodicos System which participated two years ago to the Question Answering CLEF evaluation campaign for French. The ProdicosAV system is composed of four modules whose some of them come from the PRODICOS system. We present, in this paper, the adaptation of these four modules and the new validation module. The validation module is divided into two steps : a temporal validation (comparison between the temporal elements of question and the temporal elements of passage) and a answer validation (comparison between Prodicos answer and Passage's answer). Our system obtains a precision rate equal to 0.56 and a recall rate equal to 0.46 and the qa-accuracy rate obtained is 22%.

Answer Validation on English and Romanian Languages

Adrian Iftene[1] 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

The present article presents the steps involved in the transformation of the TE system that was used in the RTE3 competition in 2007 for the AVE 2008 exercise.

This year, for our second participation in the AVE competition, we improved the system we used last year and, additionally, introduced a question analysis part, which is specific to a question answering system. In this year’s AVE competition, we also participated with a system working in Romanian, using a Textual Entailment (TE) system working on Romanian. The latter is similar to the TE system working in English with which we participated in the RTE 3 competition in 2007.

The main architecture of our Textual Entailment system remains the same. The goal of the system is to transform the hypothesis, making use of extensive semantic knowledge from resources like DIRT, WordNet, Wikipedia, and database of acronyms. The processing tools we used are LingPipe and MINIPAR. The system uses two rules in order to calculate the global fitness (representing the distance between the text and the hypothesis), namely: the Semantic Variability Rule, and the Rule for Named Entities.

The main change performed this year is regarding the Rule for Named Entities. Additionally to the system built for AVE 2007, we perform the following steps:

- Identify the Answer Type (AT) for the answers;
- Identify the Expected Answer Type (EAT) for the questions.
- Verify if AT is equal to EAT.

We submitted two runs on each of the languages (English and Romanian), according to the use or not of some system components. The systems are similar and only the external resources used by the TE system or by GATE are language-specific. The first run: is based only on TE System output. The second run: in addition to the first run, we add the comparison between EAT and AT.

The improvement in the QA accuracy obtained this year is 3% better than our results obtained in CLEF in 2007.

To conclude, we showed how changing some of the rules employed in the Textual Entailment system and adding the question and answer type classification and matching component, we improved, on the one hand, the correct classification of the answers, and on the other hand, the validation of more answers.

A Lexical-Semantic Approach to AVE

Óscar Ferrández, Rafael Muñoz and Manuel Palomar
Natural Language Processing and Information Systems Group,
Department of Computing Languages and Systems, University of
Alicante, Spain
{ofe, rafael, mpalomar}@dlsi.ua.es

In our participation in AVE 2008, we present and discuss a system capable of detecting when answers for specific questions are supported by snippets, all provided by Question Answering (QA) systems. The system uses a set of regular expressions in order to join the question and the answer into an affirmative sentence and afterwards applies several lexical-semantic inferences to attempt to detect whether the meaning of this sentence can be inferred by the meaning of the supporting text. We built a system base which consists of the computation of some lexical-similarity measures, and in order to enrich the semantic knowledge of the system we added two constraints based on Named Entities and verbs relations. These constraints are computed prior to the calculation of the lexical-similarity measures, and for this purpose we used resources such as WordNet, VerbNet, VerbOcean and an open-domain Named Entity recognizer. Moreover, we should like to apply special emphasis to the language-independent capabilities of some of our system components. As a result, we are able to apply our techniques over both Spanish and English corpora, obtaining an f-measure rate (over the positive pairs) of 0.44 and 0.49 respectively.

Justification of Answers by Verification of Dependency Relations – The French AVE Task

Véronique Moriceau, Xavier Tannier, Arnaud Grappy and Brigitte Grau
LIR Group – LIMSI (CNRS)
first_name.last_name@limsi.fr

This paper presents LIMSI results in Answer Validation Exercise (AVE) 2008 for French. In this task, systems have to consider triplets (question, answer, supporting text) and decide whether the answer to the question is correct and supported or not according to the given supporting text.

We tested two approaches during this campaign:

- A syntax-based strategy, where the system decides whether the supporting text is a reformulation of the question.

- A machine learning strategy, where several features are combined in order to validate answers: presence of common words in the question and in the text, word distance, etc.

The first system, called FIDJI, uses a syntactic parser on questions and provided passages. The approach is to detect, for a given tuple question/answer/supporting text, if all the characteristics of the question can be retrieved in the text. As in other works, some rewriting rules have been set up in order to account for syntactic variations such as passive/active voice, nominalization of verbs, appositions, coordinations, etc. Documents are also tagged with named entity types; Combined with the analysis of the question, this can be used to check that the answer corresponds to the expected type. A few heuristics are then applied to validate the answer.

The second strategy follows a machine learning approach and applies the question-answering system FRASQUES in order to compute some of the learning features. The learning set is extracted from the data provided by AVE 2006 and contains 75% of the total data. The chosen classifier is a combination of decision trees with the bagging method. It is provided by the WEKA program that allows to test a lot of classifiers. Features are terms in common between the passage and the answer (and especially the focus (main word), the answer type, the main verb and bi-terms), the answer given by our existing system FRASQUES, the longest common chain of words, the answer type checking with Wikipedia, as well as answers given by FIDJI system.

The first system leads to a very good precision (88%) but a quite low recall (42%), while the second one improves recall and reaches a F-measure of 61%.

These results must be put into perspective because of the low number of answers, and especially positive answers, provided by AVE for French this year.

Information Synthesis for Answer Validation

Rui Wang[1] and Günter Neumann[2]
[1] Saarland University
66123 Saarbrücken, Germany
rwang@coli.uni-sb.de
[2] LT-Lab, DFKI
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
neumann@dfki.de

This report is about our participation in the Answer Validation Exercise (AVE2008). Our system casts the AVE task into a Recognizing Textual Entailment (RTE) problem and uses an existing RTE system to validate answers. Additional information from named-entity (NE) recognizer, question analysis component, and so on, is also considered as assistances to make the final decision. In all, we have submitted two runs, one run for English and the other for German. They have achieved f-measures of 0.64 and 0.61 respectively. Compared with our system last year, which purely depends on the output of the RTE system, the extra information does show its effectiveness.

University of Hagen at CLEF 2008: Answer Validation Exercise

Ingo Glöckner
Intelligent Information and Communication Systems (IICS)}
University of Hagen (FernUniversitaet in Hagen)
58084 Hagen, Germany
ingo.gloeckner@fernuni-hagen.de

RAVE (Real-time Answer Validation Engine) is a logic-based answer validator/selector designed for application in real-time question answering. RAVE uses the same toolchain for deep linguistic analysis and the same background knowledge as its predecessor (MAVE), which took part in the AVE 2007. However, a full logical answer check as in MAVE was not considered suitable for real-time answer validation since it requires parsing of all answer candidates. Therefore RAVE uses a simplified validation model where the prover only checks if the support passage contains a correct answer at all. This move from logic-based answer validation to logical validation of supporting snippets permits RAVE to avoid any parsing of answers, i.e. the system only needs a parse of the question and pre-computed snippet analyses. In this way very low validation/selection times can be achieved. Machine learning is used for assigning local validation scores using both logic-based and shallow features. The resulting local validation scores are improved by aggregation. One of the key features of RAVE is its innovative aggregation model, which is robust against duplicated information in the support passages. In this model, the effect of aggregation is controlled by the lexical diversity of the support passages for a given answer. If the support passages have no terms in common, then the aggregation has maximal effect and the passages are treated as providing independent evidence. Repetition of a support passage, by contrast, has no effect on the results of aggregation at all. In order to obtain a richer basis for aggregation, an active validation approach was chosen, i.e. the original pool of support passages in the AVE 2008 test set was enhanced by retrieving additional support passages from the CLEF corpora. This technique already proved effective in the AVE 2007. The development of RAVE is not finished yet, but the system already achieved an F-score of 0.39 and a selection rate of 0.61 compared to optimal selection. Judging from last year's runs of MAVE (with a 0.93 selection rate and F-score of 0.72), this may look disappointing. However, the AVE task for German was much more difficult this year, and the F-score gain of RAVE (over the 100% yes baseline) and qa-accuracy gain (compared to random selection) are better than in last year's runs of MAVE.

INAOE at QA@CLEF 2008: Evaluating Answer Validation in Spanish Question Answering

Alberto Téllez-Valero, Antonio Juárez-González, Manuel Montes-y-Gómez and Luis Villaseñor-Pineda
Laboratorio de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrrique Erro no. 1, Sta. María Tonantzintla, Pue.; 72840; México
{albertotellezv, antjug, mmontesg, villaseng}@inaoep.mx

This paper introduces the new INAOE's answer validation method. This method is based on supervised learning approach that uses a set of attributes that capture some lexical-syntactic relations among the question, the answer and the given support text. In addition, the paper describes the evaluation of the proposed method at both the Spanish Answer validation Exercise (AVE 2008) and the Spanish Question Answering Main Task (QA 2008). The evaluation objectives were twofold. One the one hand, evaluate the ability of our answer validation method to discriminate correct from incorrect answers, and on the other hand, measure the impact of including an answer validation module in our QA system. The evaluation results were encouraging; the proposed method achieved a 0.39 F-measure in the detection of correct answers, out-performing the baseline result of the AVE 2008 task by more than 100%. It also enhanced the performance of our QA system, showing a gain in accuracy of 22% for answering factoid questions. Furthermore, when there were evaluated three candidate answers per question, the answer validation method increased the MRR of our QA system by 40%, reaching a MRR of 0.28.

The Contribution of FaMAF at QA@CLEF 2008 Answer Validation Exercise

Julio J. Castillo
Faculty of Mathematics Astronomy and Physics
National University of Cordoba, Argentina
cj@famaf.unc.edu.ar

This work is about the participation of FaMAF at AVE Challenge. The main objectives is to determinate if the answer to a question is correct or not. The inputs for the AVE systems are a set of triplets (Question, Answer, Supporting Text) and the results are a boolean value indicating whether the answer is supported by the text.

We have developed a system that performs morphological analysis (stemming, and POS tagging), and extract lexical features, and semantic features to build a model using a Support Vector Machine (SVM), to determinate whether the implication is holds. The system utilizes a Recognizing Textual Entailment (RTE) approach. The main idea is to think in the question string (q_str) as a Text (T) and one answer (t_str) as a Hypothesis (H).

We have developed two experiments. The first experiment (Run 1) consists of twelve lexical features, and the other (Run 2) take in account only three features (lexical and semantic). The Run 2 obtains best results that Run 1. The features used are unigram, bigram, and trigram overlap of lexemes and stems, Levenshtein distance, tf-idf measure, and semantic similarity using Wordnet. We took all TRUE pairs from the training sets in AVE 2006 and AVE 2007 and then we incorporated a number of FALSE pairs totalling a 40% of the total. We intended that the system could characterize in special the true answers.

The official results shows that Run 1 obtains an F-measure of 0.17, precision of 0.09, recall of 0.94, and QA accuracy of 0.16. On the other hand, Run 2 obtains an F-measure of 0.21, precision of 0.13, recall of 0.56, and QA accuracy of 0.17. In spite of the simplicity of the approach, we have obtained a reasonable 0.17 of QA accuracy for our best Run (Run 2).

The results show an increment over the baselines, however enhanced is needed.

Future work is oriented to probe with different classifiers as Bayesian Binary Regression (BBR), and use different datasets RTE, and RTE+AVE. To enhance the system, we will work with lexical and semantic similarity, adding features and testing his improvement. Additionally an NER module will be incorporated and combined with the rest of the system and his performance will be evaluated.

Question Answering on Speech Transcription (QAST)

QA Extension for Xtrieval: Contribution to the QAst track

Jens Kürsten, Holger Kundisch and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media
09107 Chemnitz, Germany
[jens.kuersten j holger.kundisch j maximilian.eibl] at cs.tu-chemnitz.de

This article describes our first participation at the QAst task of the CLEF campaign 2008. We submitted 4 experiments in total, two for each subtask t1 and t4. These subtasks employed manual speech transcription collections. Our main goal was to implement a QA prototype that is able to answer posed questions with a high accuracy and a acceptable recall. We used the Stanford Named Entity Recognizer for tagging named entities and the CRFTagger - Conditional Random Fields Part-of-Speech (POS) Tagger for English. The passage retrieval was done with the Xtrieval framework and its Apache Lucene implementation. For the classification of the question hand-crafted patterns were implemented. Our experiments achieved an accuracy of about 20%, which meets our expectations. Although, the rate of returned NIL answers was too high for all of our experiments. The participation at the QAst task helped us to identify the main problems of a QA system and inspired us to some ideas for further improvement of the system.

The LIMSI Participation to the QAst Track

Sophie Rosset, Olivier Galibert, Guillaume Bernard, Eric Bilinski and Gilles Adda
LIMSI-CNRS, France

LIMSI participated to the QAST 2008 evaluation. This year 5 document types were used: meetings and seminars in English, broadcast news in French and European Parliament Plenary Sessions in English and Spanish. For each document type a manual transcription and up to three automatic transcriptions with varied error rates were provided, for a total of 16 subtasks. We participated to all the subtasks and submitted 18 runs (2 runs for two of the subtasks). The evaluation results ranged from 31% to 45% for accuracy for manual transcriptions and from 16 to 41% for automatic transcripts.

Our system this year is a refinement of last year's. Our efforts went essentially into the new document types and languages and on the general robustness of the system. The structure stayed the same: a language-dependant analysis common to questions and document followed by a language-independant search engine/answer extractor combo.

The analysis tries to find the bits of information that are useful for the search and extraction. They are of different categories: named entities, linguistic entities (e.g. verbs, prepositions), or specific entities (e.g. scores). The French analyser detects around 300 types and constitutes the basis for the Spanish and English EPPS analysers. The Spanish analyser is a simple adaptation of the French one with only a lexicon-level adaptation. English required a deeper adaptation, in particular the order in which the blocks of rules are applied is reversed. The English and Spanish analysers detect only about a hundred types.

Once the question is analysed a Search Descriptor is built which contains the elements of the input considered pertinent for the search and the expected type or types for the answer. The documents are scored using the counts of occurrences of the SD elements, ponderated by the SD weights. Snippets are then extracted from them based on the presence of the SD elements, and scored similarly. Every element in a snippet of one of the expected answer types of the SD is considered an answer candidate and is given a score, based on the distance between itself and the elements of the SD and its redundancy. Additionally this year we tried a new experimental re-scoring method based on tree transformation costs which justified the two extra runs.

Adapting IBQAS to Work with Text Transciptions in QAst Task: IBQAst

M. Pardiño, J.M. Gómez, H. Llorens, R. Muñoz-Terol, B. Navarro-Colorado, E. Saquete, P. Martínez-Barco, P. Moreda and M. Palomar
Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Natural Language Processing and Information Systems Group Department of Software and Computing Systems
University of Alicante, Spain
maria,jmgomez,hllorens,rafamt,borja,stela,patricio,moreda,mpalomar@dlsi.ua.es

The paper shows the results of adapting a modular domain English QA system (called IBQAS, whose initials correspond to Interchangeable Blocks Question Answering System) to work with both manual and automatic text transcriptions. This system provides a generic and modular framework using an approach based on the recognition of named entities as a method of extracting answers. The system architecture follows the general methodology of QA systems incorporating the modules detailed below: analysis of the question, information retrieval and extraction of the answer. In the analysis phase of the system, we extracted the type of question or type of answer expected, keywords and focus. Next, we used JIRS, a traditional Passage Retrieval system which is able to find structures in questions using n-gram models, for the information retrieval process. Finally, we selected the potential answers and those with higher scores were given as result. The best results have been obtained with the manual transcription. This is due to the fact that this transcription has fewer errors than automatic transcriptions because most of the problems have been checked manually.

Robust Question Answering for Speech Transcripts: UPC Experience in QAst 2008

Pere R. Comas and Jordi Turmo
TALP Research Center, Technical University of Catalonia (UPC)
{pcomas,turmo}@lsi.upc.edu

This paper describes the participation of the Technical University of Catalonia in the CLEF 2008 Question Answering on Speech Transcripts track. We have participated in all tasks of the English and Spanish scenarios of QAst obtaining the best results in some of these tasks. For some tasks, different word error rate (WER) automatic transcripts exists, allowing detailed analysis of ASR effect on QA.

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 Passage Retrieval and Answer Extraction engine based on a sequence alignment algorithm that searches for ``sounds like'' sequences. We use a machine-learning based NERC enhanced with phonetic features for automatic transcripts.

Our approximated keyword search algorithm used for passage retrieval obtains mixed results. It can improve standard search for Spanish but makes little difference for English. We think this because in some document collections it may generated too many false-positive, introducing noise in sets of candidate passages and answers.

We perform a detailed analysis of our results and it shows that automatic speech recognition has critical impact on the performance of NERC but its affect on passage retrieval is much less severe. The performance of the NERC decreases exponentially as WER increases, but passage retrieval decreases only linearly with WER.

QA and Word Sense Disambiguation (QA-WSD)

QA with a Disambiguated Document Collection

Davide Buscaldi and Paolo Rosso
Natural Language Engineering Lab,
Dpto. Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia
{dbuscaldi, prosso}@dsic.upv.es

In this report we describe our approach to the Question Answering - Word Sense Disambiguation task. In our approach, disambiguated documents are used to improve the retrieval phase: this has been implemented by adding a WordNet expanded index to the document collection. This index contains synonyms, hyperonyms and holonyms of the words contained in the documents. Question words are searched for in both the expanded WordNet index and the default index. The obtained results do not show any improvement over the system that do not use the disambiguated collection. However, an analysis of the results shows that the average number of passages that contains the answer for each question (2.04) is too small to be used to discriminate between the two systems. Another issue is that the question set was not dedicated to this task, with about 75% of the questions having their answers in Wikipedia and not in the disambiguated collection.

Exploiting Cooccurrence on Corpus and Document Level for Fair Crosslanguage Retrieval

Andreas Juffinger, Roman Kern and Michael Granitzer
Know-Center, Graz
ajuffinger,rkern,mgranitzer@know-center.at

In our work: exploiting cooccurrence on corpus and document level for fair crosslanguage retrieval, for the Robust WSD Task in the Question Answering Track at CLEF2008, we have developed a text retrieval system which is based on extensive query preprocessing, followed by standard text retrieval techniques. The preprocessing of queries includes: An optional query expansion step based on Wordnet Synonyms or an Associative Index. The wordnet query expansion strategy makes use of the available word sense disambiguation information. For each term that is tagged with a Wordnet senses the sense with highest score is selected, from which we then take the appropriate synonyms as query expansion terms. For the associative query expansion the query terms are used as start nodes in the term network, calculated from the cooccurrence term statistics. The top most adjacency terms from this network were then included in the query, restricted by the maximum number of terms and minimal similarity value. The associative network is constructed by exploiting the term cooccurrences within the documents of a corpus. The algorithm to build this network uses a sliding window approach to calculate the weights between cooccurring terms within certain term vicinity. For this task we have build this associative index separately for English and Spanish on our multilingual Wikipedia index. We construct the multilingual Wikipedia index from the Spanish and English Wikipedia XML dumps. Additionally the multilingual index contains links from the Spanish articles to the same articles in the English Wikipedia. In the query translation step, that follows the query expansion, each query term forms a separate query in the multilingual index. For English terms we query the English Wikipedia articles and for Spanish the Spanish ones. By exploiting the crosslanguage links within the multilingual index we collect the top 50 documents in the language of the corpus for which the preprocessing is done, for this task the CLEF corpus. These documents are then used to extract the final query terms to construct a disjunct index search in the CLEF corpus.

The system applies all steps to all query-corpus language pairs and ensures therefore fairness across different languages. We have been able to show that our query translation technique allows crosslingual retrieval with minor impact in monolingual retrieval - the fairness comes therefore at nearly no cost. We agree totally with the organizers of this task, that WSD information should have a significant positive impact in the retrieval task. Unfortunately we have not been able to show a significant improvement in the performance when including word sense disambiguation information in our setup.

Cross-Language Retrieval in Image Collections

Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task

Thomas Arni[1], Paul Clough[1], Mark Sanderson[1] and Michael Grubinger[2]
[1] Sheffield University, Sheffield, UK
[2] Victoria University, Melbourne, Australia

ImageCLEFphoto 2008[1] is an ad-hoc photo retrieval task and part of the ImageCLEF evaluation campaign. This task provides both the resources and the framework necessary to perform comparative laboratory-style evaluation of visual information retrieval systems. In 2008, the evaluation task concentrated on promoting diversity within the top 20 results from a multilingual image collection, which has been shown to better satisfy a user's information need. Similar to the 2006 and 2007 ImageCLEFphoto tasks, we generated a subset of the IAPR TC-12 Benchmark as an evaluation resource for 2008: 20,000 colour photographs and two sets of semistructured annotations in (1) English and (2) one set whereby the annotation language was randomly selected from English and German for each of the images. From an existing set of 60 topics, 39 were selected and distributed to participants representing varying search requests and suitable for evaluating diversity.

To enable diversity to be quantified, it was necessary to classify images relevant to a given topic to one or more sub-topics or clusters [1]. This was performed by two assessors. In case of inconsistent judgements, a third assessor was used to resolve the inconsistencies. The resulting cluster assessment judgements were used in combination with the normal relevance assessment to determine the retrieval effectiveness of each submitted system run. The results for submitted runs were computed using the latest version of trec eval, as well as a custom-built tool to calculate diversity of the results set. Submissions were evaluated using two metrics: (1) precision at rank 20 (P20) and (2) cluster recall at rank 20 (CR20). To enable absolute comparison between individual runs, the F1-measure was used to combine scores from P20 and CR20 (representing the harmonic mean of P20 and CR20).

This new challenge attracted a record number of submissions: a total of 24 participating groups submitting 1,042 system runs. These were categorised with respect to the following dimensions: (1) annotation language, (2) modality (text only, image only or combined) and (3) run type (automatic or manual). Most submissions (96.8%) used the provided image annotations, with 22 groups submitting a total of 404 purely concept-based (textual) runs and 19 groups a total of 605 runs using a combination of content-based (visual) and concept-based features. A total of 11 groups submitted 33 purely content-based runs. Of all retrieval approaches, 61.2% involved the use of image retrieval (53.4% in 2007 and 31% in 2006), 79% of all groups used content-based (i.e. visual) information in their runs (60% in 2007 and 58% in 2006). Almost all of the runs (99.7%) were automatic (i.e. involving no human intervention); only 3 submitted runs were manual.

Some of the findings include that the choice of annotation language is almost negligible (the best performing runs using random annotations performed with an F1-measure score at 97.4% of the highest monolingual run) and the best performing runs combine concept and content-based retrieval methods in a two-stage process: standard ad-hoc retrieval followed by some form of clustering to promote diversity in the top 20 results.

[1] Arni, T., Tang, J., Sanderson, M. and Clough, P. (2008) Creating a test collection to evaluate diversity in image retrieval, In Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments, held at SIGIR2008.

Overview of the ImageCLEFmed 2008 Medical Image Retrieval Task

Henning Müller[1,2], Jayashree Kalpathy-Cramer[3], Charles E. Kahn Jr. [4], William Hatt[3], Steven Bedrick[3] and William Hersh[3]
[1]Medical Informatics, University Hospitals and University of Geneva, Switzerland,
[2]University of Applied Sciences Western Switzerland, Sierre, Switzerland,
[3]Oregon Health and Science University (OHSU), Portland, OR, USA,
[4]Department of Radiology, Medical College of Wisconsin, Milwaukee, WI, USA
henning.mueller@sim.hcuge.ch

2008 was the fifth year for the medical image retrieval task of ImageCLEF, one of the most popular tracks within CLEF. Registration continued to increase in 2008. A total of 15 groups submitted 111 valid runs. Several requests for data access were also received after the registration deadline.

The most significant change in 2008 was the use of a new database containing images from the biomedical literature. These images, part of the Goldminer collection, were from the journals Radiology and Radiographics. Besides the images, the figure captions and the part of the caption referring to a particular sub figure were supplied to the participants. Access to the full text articles in HTML was also provided, as was each article's Medline PMID (PubMed Identifier). An article's PMID could be used to obtain the officially assigned MeSH (Medical Subject Headings) terms. Unlike previous years, this year's collection was entirely in English, as it was obtained from English-language medical literature. However, the topics were, as in previous years, supplied in German, French, and English. The topics used in 2008 were a subset of the 85 topics used in 2005-2007. Thirty topics were made available, ten in each of three categories: visual, mixed, and semantic.

As in previous years, most groups concentrated on fully automatic retrieval. However, three groups submitted a total of six manual or interactive runs; these runs did not show a substantial increase in performance over the automatic approaches. In previous years, multi-modal combinations were the most frequent submissions. However, in 2008 only half as many mixed runs as purely textual runs were submitted. Very few fully visual runs were submitted, and the ones submitted performed poorly. This may be explained in part by the heavily semantic nature of the 2008 topics.

From examining mixed media runs that had corresponding text-only runs, it is particularly clear that combining good textual retrieval techniques with questionable visual retrieval techniques can negatively affect system performance. This demonstrates the difficulty of usefully integrating both textual and visual information, and the fragility that such combinations can introduce into retrieval systems.

The best MAP scores were very similar for textual and multi--modal approaches, whereas early precision performance was clearly better for the multi-modal approaches.

Medical Image Annotation in ImageCLEF 2008

Thomas Deselaers[1] and Thomas M. Deserno[2]
[1]RWTH Aachen University, Computer Science Department, Aachen, Germany
[2]RWTH Aachen University, Dept. of Medical Informatics, Aachen, Germany
deselaers@cs.rwth-aachen.de, deserno@ieee.org

The ImageCLEF 2008 medical image annotation task is designed to assess the quality of content-based image retrieval and image classification by means of global signatures. In total, 12,076 images were used. In contrast to previous years, the task was designed such that the hierarchy of reference IRMA code classifications is essential for good performance. 24 runs of 6 groups were submitted. Multi-class classification schemes for support vector machines outperformed the other methods. The obtained scores rage from 74.92 over 182.77 to 313.01 for best, baseline and worst results, respectively.

The Visual Concept Detection Task in ImageCLEF 2008

Thomas Deselaers[1] and Allan Hanbury[2]
[1]RWTH Aachen University, Computer Science Department, Aachen, Germany
[2]PRIP, Institute of Computer-Aided Automation, Vienna University of Technology, Austria
deselaers@cs.rwth-aachen.de, hanbury@prip.tuwien.ac.at

The Visual Concept Detection Task (VCDT) of ImageCLEF 2008 is described. A database of 2,827 images were manually annotated with 17 concepts. Of these, 1,827 were used for training and 1,000 for testing the automated assignment of categories. In total 11 groups participated and submitted 53 runs. The runs were evaluated using ROC curves, from which the Area Under the Curve (AUC) and Equal Error Rate (EER) were calculated. For each concept, the best runs obtained an AUC of 80% or above.

Overview of the WikipediaMM Task at ImageCLEF 2008

Theodora Tsikrika[1] and Jana Kludas[2]
[1] CWI, Amsterdam, The Netherlands
[2] CUI, University of Geneva, Switzerland
Theodora.Tsikrika@cwi.nl, jana.kludas@cui.unige.ch

ImageCLEF's wikipediaMM task provides a testbed for the system-oriented evaluation of ad hoc multimedia retrieval from a collection of Wikipedia images. The aim is to investigate mono-media and cross-media retrieval approaches in the context of a larger scale and heterogeneous collection of images (similar to those encountered on the Web) that are searched for by users with diverse information needs. In 2008, ImageCLEF wikipediaMM used the collection of Wikipedia images previously used in the INEX 2006-2007 Multimedia evaluation campaign. This collection contains approximately 150,000 images on diverse topics associated with unstructured and noisy textual annotations in English. The following resources were also made available to the participants: (i) low level features extracted from the images, and (ii) classification scores for the 101 MediaMill visual concepts computed using the extracted features.

The topics are descriptions of multimedia information needs that contain a textual query and optional visual evidence in the form of image examples and visual concepts. This year's topics were developed in cooperation with the participants. The final set contains 75 topics: 5 visual, 35 textual and 35 semantic. The assessments were also performed by the participants, who assessed from each topic the set of top 100 retrieved images pooled from all submitted runs.

A total of 12 participating groups submitted 77 runs to the wikipediaMM task. While many of the runs are textual only approaches (45%), there is also a significant trend towards fusion approaches that combine evidence from different modalities: text and images (29%), text and concepts (26%), and text, images, and concepts (3%). The main evaluation measure was Mean Average Precision (MAP). Additional measures include precision at 20 and R-precision. Our analysis of the average performance per modality indicates that the runs that fuse textual and conceptual evidence achieve a higher MAP (0.2316) than the textual-only runs (0.2137). This result is supported by our per topic analysis that indicates that approaches combining text and concepts outperform the text-only ones for 62% of the topics. So, we can conclude that multi-modal fusion approaches can help to improve the retrieval performance in this domain.

ImageCLEFphoto

LIG at ImageCLEFphoto 2008

Philippe Mulhem
UJF, UMR CNRS 5217, Laboratoire d'informatique de Grenoble
Philippe.Mulhem@imag.fr

This abstract describes the runs and results obtained by the LIG at ImageCLEFphoto 2008. The submitted runs are: two runs (text only and text+image) without diversification on classes, and two runs (text only and text+image) with class diversification were submitted. The main idea on this task was to make a first step in using language models for both text and image, even if for the work done the image part is not fully a language model. The text retrieval is based on language model of Information Retrieval, and the image part is processed using RGB histograms on 9 image blocks with a similarity value based on Jeffrey divergence. Results using text+image are obtained by a linear combination of normalized results on text and image. The diversification is based on clusters, according to the cluster given in the queries. When the cluster name is not directly extracted from the images (like city or country), we apply a visual clustering. All the processes (text and image processing, as well as retrieval), were done on a Linux computer (dual-core). Not surprisingly, the cluster recall at 20 (i.e., cr(20)) results are higher for the runs that include diversification. On the other hand, the precision at 20 and the mean average precision results are higher without diversification on our runs, for both text only and image+text results.

Meiji University at ImageCLEF2008 Photo Retrieval Task: Evaluation of Image Retrieval Methods Integrating Different Media

Kosuke Yamauchi, Takuya Nomura, Keiko Usui, Yusuke Kamoi, Miki Eto and Tomohiro Takagi
Meiji University, Japan
Email: {yamauchi, takagi} @cs.meiji.ac.jp

This paper describes the participation of the Human Interface Laboratory of Meiji University in the ImageCLEF2008 photo retrieval task.

We submitted eight retrieval runs taking two main approaches. The first approach combined Text-Based Image Retrieval (TBIR) and Context-Based Image Retrieval (CBIR). We used Apache Lucene as the TBIR, and CBIR consists of three retrieval modules called global retrieval, grid retrieval, and region retrieval. We expect that each fault can be avoided by combining TBIR and CBIR, and the accuracy of the image retrieval can be improved by the synergistic effect of different media used to solve these problems. The second approach applied query expansion using conceptual fuzzy sets (CFS). CFS is a method that uses the expression of meaning depending on the context, which an ordinary fuzzy set does not recognize. A conceptual dictionary is necessary to perform query expansion using CFS and this is constructed by clustering. We propose a system that depends more on query context than on query expansion for improving the packaging method of a CFS. We propose here the use of query expansion with CFS, and other techniques, for image retrieval that integrates different media, and we verify the utility of the system by explaining our experimental results.

As for result, it was understood that integrating different media, i.e., TBIR and CBIR, was conducive to the higher retrieval result. Moreover, when looking at the mean average precision (MAP), the top three systems perform query expansion using CFS. This indicates that performing query expansion using CFS produces a higher retrieval result.

We demonstrated that retrieval accuracy improved by performing query expansion using CFS in image retrieval that integrates different media. However, future tasks remain, and these are as follows. It is necessary to improve the accuracy because the accuracy of CBIR is low. Also, another problem is how to determine the initial point of clustering when the conceptual dictionary is constructed. It is thought that accuracy will fall outside the accuracy range if another corpus is used, but there is the possibility that accuracy will be higher or lower than the results reported here when the conceptual dictionary is constructed anew because this system sets the initial point at random.

Building a Diversity Featured Search System by Fusing Existing Tools

Jiayu Tang, Thomas Arni, Mark Sanderson and Paul Clough
Department of Information Studies, University of Sheffield, UK
{j.tang, t.arni, m.sanderson, p.d.clough}@sheffield.ac.uk

We have participated in the task of ImageCLEFPhoto 2008. The main objectives of our experiments are to build a diversity featured retrieval system and tune the system in order to achieve better performance. Two existing tools are used: Solr and Carrot2. Solr is a text search server and Carrot2 is a search results clustering engine. We have changed 4 kinds of settings: the field used for clustering, the number of images used for clustering, indexing and query expansion, and parameters of the clustering algorithm. The results suggest that indexing and query expansion can fairly improve precision. Moreover, appropriately chosen clustering method can increase diversity of the results while keeping precision almost the same.

TELECOM ParisTech at ImageClefphoto 2008: Bi-Modal Text and Image Retrieval with Diversity Enhancement

Marin Ferecatu and Hichem Sahbi
Institut TELECOM, TELECOM ParisTech
CNRS LTCI, UMR 5141
46, rue Barrault, 75634 Paris Cedex, France
Email: Marin.Ferecatu@telecom-paristech.fr, Hichem.Sahbi@telecom-paristech.fr

We describe here the participation of TELECOM ParisTech in the ImageClefphoto 2008 challenge. This edition focuses on promoting diversity in the results produced by the retrieval systems. Given the high level semantic content of the topics, search engines based solely on text or visual descriptors are unlikely to offer satisfactory results. Our system uses several text and visual descriptors, as well as several combination algorithms to improve the overall retrieval performance. The text part includes a collection of manually built boolean queries and a set of textual descriptors extracted automatically using dictionary filtering and dimensionality reduction. Text and visual descriptors are combined using two strategies: ad-hoc concatenation and re-ranking. Diversity makes it possible to reduce the redundancy in the final results and it is obtained using two techniques, threshold clustering and Maxmin exploration. Several runs were submitted to the challenge, including individual (text or visual), combined, and with different settings of diversity.

First, we found that even with very few images (three in the ImageClefphoto), the combined runs outperform by a significant amount the individual runs. Moreover, the improvement is more significant in case of manually prepared boolean queries, where our runs were ranked 1st, 2nd and 3rd. This clearly indicates that good quality boolean queries are less likely to return noisy results with respect to the targeted topic. Automatic extraction of boolean queries from raw text is hence identified as a worthy to explore research direction, for instance by using Parts of Speech (POS) tagging and language parsing.

Second, using a diversification algorithm definitely improved the ranking of our runs. This is more noticeable for queries using only visual descriptors where the proposed diversification schemes significantly improved the ranking of our runs (2-nd and 3-rd place). However, because of the limited size of ground truth classes (less than 100 images per topic), it is not possible to draw firm conclusions. Indeed, in a real search engine, where topics might be represented by millions of (possibly similar) images, we expect the obtained clusters to be much more consistent.

Affinity Propagation Promoting Diversity in Visuo-Entropic and Text Features for CLEF Photo Retrieval 2008 Campaign

Herve Glotin and Zhongqiu Zhao
Laboratoire des sciences de l'information et des systemes
UMR CNRS & Université Sud Toulon-Var France
glotin@univ-tln.fr, zhongqiuzhao@gmail.com

We develop for the CLEF PHOTO 2008 task a new visual feature using various pixel projections for training SVMs, allowing to produce image retrieval and clustering using affinity propagation. To heighten the diversity of the top of the retrieval results, we put the images with the lowest rank in each cluster into the top. The LSIS run which used only the visual information is at the 6th best team rank in the AUTO IMG run type. For AUTO TXTIMG runs, we merge by simple harmonic or arithmetic average our visual ranks to the textual ranks of the LIG language model participating to the AVEIR consortium. Then we also perform the affinity propagation and the re-ranking on this TXTIMG run, which gives complementary information to the AVEIR consortium, helping in producing the third best AUTO TXTIMG run (after XEROX). We discuss on the clustering performance of the various run types, and then we give some perspectives for enhancing such diversity image retrieval system. If affinity propagation clustering seems efficient for promoting visual diversity, our results show that clustering process itself should merge independent textual and visual clustering information.

MIRACLE-FI at ImageCLEFphoto 2008: Experiences in Merging Text-Based and Content-Based Retrievals

R. Granados[1], X. Benavent[2], A. García-Serrano[3] and J.M. Goñi-Menoyo[1]
1 Universidad Politécnica de Madrid
2 Universidad de Valencia
3 Universidad Nacional de Educación a Distancia, UNED
{rgranados@fi.upm.es, xaro.benavent@uv.es, agarcia@lsi.uned.es, josemiguel.goni@upm.es}

This paper describes the participation of the MIRACLE consortium at the ImageCLEF Photographic Retrieval task of ImageCLEF 2008. In this new participation our first purpose with our experiments is to evaluate our own tools for text-based retrieval and for content-based retrieval using different similarity metrics and the aggregation OWA operator to the three topic images.

From the MIRACLE last year experience, we implemented a new merging module combining the text-based and the content-based information in three different ways: FILTER-N, ENRICH and TEXT-FILTER. The former approaches try to improve the text-based baseline results using the content-based results lists. The last one was used to select the relevant images to the content-based module. No clustering strategies were analyzed.

Selected fields from images annotations taken into account to index are: TITLE, DESCRIPTION, NOTES and LOCATION. In the case of the queries, TITLE and NARR fields are selected. The visual retrieval module uses color and texture descriptors for extracting the features of the images; then an OWA aggregation operator is used to combine the three topic images, and finally, a similarity distance is calculated for ranking the images of the database.

Finally, 41 runs were submitted: 1 for the text-based baseline, 10 content-based runs, and 30 mixed experiments merging text and content-based results. Results in general can be considered nearly acceptable comparing with the best results of other groups.

Obtained results with the textual-based retrieval module can be considered acceptable, having into account that no linguistic processes were applied. The MAP (0.2253) is higher than the average MAP taken from the best 4 runs for each participating group (0.2187).

For the content-based image module was testing we can observe that the Mahalanobis distance outperforms the Euclidean distance, and the best aggregation method in both metrics is the minimun (AND), followed by the orness(W)_0.3 that is a smoothed AND. Our best result for this group of experiments is the combination of the Mahalanobis metrics with orness(W)_0.3 with a MAP(0.0213) and a P20(0.0679). Our best result is considerably lower than the best result for this group.

The FILTER-10000 merge algorithm improves the baseline in the precision at low values (5, 10) but never improves the MAP value nor the number of relevant images retrieved.

ENRICH merge method improves the baseline experiment in the MAP value and in the number of relevant images retrieved. Best MAP value (0.2401) is achieved merging the textual results with the visuals obtained using the Mahalanobis distance and the AND operator. This value is quite bigger than the average MAP taken from the best 4 runs from each participating group (0.2187).

From these results we were going to try to improve merged results by clustering methods applied to this image collection.

MIRACLE-GSI at ImageCLEFphoto 2008: Experiments on Semantic and Statistical Topic Expansion

Julio Villena-Román[1,3], Sara Lana-Serrano[2,3] and José C. 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

For the ImageCLEF Photographic Retrieval task of ImageCLEF 2008, MIRACLE team decided to split into two subgroups, MIRACLE-GSI (Grupo de Sistemas Inteligentes – Intelligent System Group) in charge of purely textual experiments, and MIRACLE-FI (Facultad de Informática, Computer Science Faculty) in charge of visual and mixed runs. This paper describes the participation of MIRACLE-GSI at ImageCLEFphoto 2008. The main purpose of our experiments for this campaign was to compare among different strategies for topic expansion in a pure textual information retrieval context. Two approaches were used: methods based on linguistic information such as thesauri, and statistical methods that use term frequency. We also participated in the ImageCLEF Medical Retrieval task with the same approach, thus allowing for comparison of results between these two different domains.

Based on our experience in previous campaigns, we designed a flexible system composed of a set of small components that may be easily combined in different configurations and executed sequentially to build the final result set. Our system is composed of five main blocks: the textual (text-based) retrieval module, which indexes image annotations in order to search and find the most relevant ones to the text of the topic; the expander module, which expands documents and/or topics with additional related terms using textual and/or statistical methods; the relevance-feedback module, which allows to execute reformulated queries that include the results of previous queries; the result combination module, which uses OR operator to combine, if necessary, the results of the previous subsystems; and, finally, a clustering module that reranks the result list to allow cluster diversity.

First a common baseline algorithm is used in all experiments to process the document collection: text extraction, tokenization, conversion to lowercase, filtering, stemming and finally, indexing and retrieval. Then this baseline algorithm is combined with different expansion techniques.

For the semantic expansion, we used (Euro)WordNet to expand topic terms with related terms corresponding to a variety of semantic relationships (mainly synonyms and hyponyms). The statistical method consisted of expanding the topics using the well-known Agrawal's apriori algorithm. This algorithm was used to discover out rules having the topic terms in their antecedents and a confidence value greater than a given threshold. Then the topic was expanded with the consequents of those (one-term) rules, i.e., with terms that are related to the topic according to the document corpus.

Additionally, relevance-feedback techniques were also used.

The last step of the process is to rerank the result list, using an implementation of k-Medoids clustering algorithm with the target number of clusters set to 20. Cluster prototypes are moved to the top positions of the final result list to allow cluster diversity and maximize cluster recall.

14 fully-automatic runs were finally submitted. In general, results are on the average, comparing to other groups. The best result in terms of MAP is achieved by the baseline experiments both for English and Random (mixed) language. However, the best cluster precision (CR), which was the variable to maximize in this task, is achieved (with a significant improvement) when k-Medoids algorithm is applied, thus proving to be valuable.

MAP values are similar in practice for experiments using topic expansion and significantly worse in the case of relevance-feedback. This shows that no strategy for either topic expansion or specially relevance-feedback has proved to be useful. The reranking algorithm used for combining the different result lists is likely to be the main reason for the disappointing results. The last conclusion that can be drawn is that the application of clustering techniques smoothes the negative effect of the expansion processes, showing quite promising results.

UPMC/LIP6 at ImageCLEFphoto 2008: on the Exploitation of Visual Concepts (VCDT)

Sabrina Tollari, Marcin Detyniecki, Ali Fakeri-Tabrizi, Massih-Reza Amini and Patrick Gallinari
Université Pierre et Marie Curie-Paris6, Laboratoire d'Informatique de Paris 6 - UMR CNRS
7606, 104 avenue du président Kennedy, 75016 Paris, France
firstname.lastname@lip6.fr

In this work, we focus our efforts on the study of how to automatically extract and exploit visual concepts. First, in the Visual Concept Detection Task (VCDT), we look at the mutual exclusion and implication relations between VCDT concepts in order to improve the automatic image annotation by Forest of Fuzzy Decision Trees (FFDTs). In our experiments, the use of relations does not improve nor worsen the quality of the annotation. Our best VCDT run is the 4th ones under 53 submitted runs (3rd team under 11 teams). Second, in the Photo Retrieval Task (ImageCLEFphoto), we use the FFDTs learn in VCDT task and WordNet to improve image retrieval. We analyse the influence of extracted visual concept models to the diversity and precision. This study shows that there is a clear improvement, in terms of precision or cluster recall at 20, when using the visual concepts explicitly appearing in the query.

Consortium AVEIR at ImageCLEFphoto 2008: on the Fusion of Runs

Sabrina Tollari[1], Marcin Detyniecki[1], Marin Ferecatu[2], Hervé Glotin[3], Philippe Mulhem[4], Massih-Reza Amini[1], Ali Fakeri-Tabrizi[1], Patrick Gallinari[1], Hichem Sahbi[2] and Zhong-Qiu Zhao[3]
[1]Université Pierre et Marie Curie-Paris6, UMR CNRS 7606-LIP6, Paris
firstname.lastname@lip6.fr
[2]TELECOM ParisTech, UMR CNRS 5141 LTCI, Paris
firstname.lastname@telecom-paristech.fr
[3]Université du Sud Toulon-Var, UMR CNRS 6168 LSIS, Toulon
name@univ-tln.fr
[4]Université Joseph Fourier, UMR CNRS 5217 LIG, Grenoble
firstname.lastname@imag.fr

We present the submission of the AVEIR consortium, composed of 4 French laboratories, to ImageCLEFphoto 2008. The submitted runs correspond to different fusion strategies applied to four individual ranks, each proposed by an AVEIR consortium partner. In particular, we study the complete, and partial, average of the ranking values, the minimum of these values, and a random based diversification. We first briefly describe the individual run of each partner, then we describe the fusion runs. The official results classed one of the runs, the MEAN fusion, as the third best in the automatic text-image run category. This run gives better results than the best partner run.

Feature Annotation for Visual Concept Detection in ImageCLEF 2008

Jingtian Jiang, Xiaoguang Rui and Nenghai Yu
Microsoft Key Laboratory of Multimedia Computing and Communication, Department of EEIS, University of Science and Technology of China
silyt@mail.ustc.edu.cn, davidrui@gmail.com, ynh@ustc.edu.cn

This paper shows our work on CLEF 2008. Our group joined the Visual Concept Detection Task of ImageCLEF 2008 this year. We submitted one run (run id: HJ_FA) for the evaluation. In the run, we applied a method called ¡°Feature Annotation¡± to detect visual concept for the predefined concepts and we want to know how this information help in solving the photographic retrieval task. The applied method selected high level features for each concept from both local and global features, based on which the visual concepts are detected. The applied method consists of three procedures. First, feature extraction in which both local and global features are extracted from images. Then, a clustering algorithm is applied to ¡°annotate the features¡±. In this procedure, the features are affiliated with their corresponding concepts. Finally, we applied KNN algorithm to classify tests images according to the training images with the annotated features. The experiments were performed on the given training and test data on the 17 concepts. The paper concludes with an analysis of our results. Finally we identify the weaknesses in our approach and ways in which the algorithm could be optimized and improved.

Targeting Diversity in Photographic Retrieval Task with Commonsense Knowledge

Supheakmungkol Sarin and Wataru Kameyama
Graduate School of Global Information and Telecommunication Studies, Waseda University
1011 Okuboyama, Nishi-Tomida, Honjo-shi, Saitama-ken 367-0035, Japan
{mungkol@fuji.waseda.jp, wataru@waseda.jp}

Image search engines have a very limited usefulness since it is still difficult to provide different users with what they are searching for. This is because most research efforts to date have only been concentrating on relevancy rather than diversity which is also a quite important factor, given that the search engine knows nothing about the user's context. In this paper, we describe our approach for ImageCLEF 2008 photographic retrieval task. The novelty of our technique is the use of AnalogySpace, the reasoning technique over commonsense knowledge for document and query expansion, which aims to increase the diversity of the results. Our proposed technique combines AnalogySpace mapping with other two mappings namely, location and full-text. We then re-rank the resulting images from the mapping by trying to eliminate duplicate and near duplicate results in the top 20. We present our preliminary experiments and the results conducted using the IAPR TC-12 photographic collection with 20,000 natural still photographs. The results show that our integrated method with AnalogySpace yields slightly better performance in terms of cluster recall and the number of relevant photographs retrieved. We finally identify the weakness in our approach and ways on how the system could be optimized and improved.

SINAI at ImageCLEFPhoto 2008

M.A. García-Cumbreras, M.C. Díaz-Galiano, M.T. Martín-Valdivia and L.A. Ureña-López
SINAI Research Group. Computer Science Department. University of Jaén
{magc,mcdiaz,maite,laurena}@ujaen.es

The SINAI system is automatic (without user interaction), and works with English text information (not visual information). The English collection documents have been preprocessed as usual (English stopwords removal and the Porter's stemmer). Then, it has been indexed using as IR systems: Lemur and Jirs.

The use of the cluster term has been oriented in a filtering way. The cluster term is expanded with its WordNet synonyms (the first sense). Then, the list of relevant documents generated by the IR system is filtered. Finally, the new list with the filtered documents is combined with the original ones (Lemur and Jirs) in order to improve them.

In general, the results in term of MAP or other precision values are not so different, with a best MAP value of 0,2125. Between the best MAP and the worse one the difference is less than 8%. Filtering methods have not improved the baseline cases. After an analysis of the performance one reason is that some relevant documents that appear in the first retrieval phase have been deleted because they not contain the cluster term. For these documents the cluster term is not useful in a filtering process. On the other hand some documents retrieved by the IR that are not relevant contain synonyms of the cluster term, so they are not deleted and the precision decrease.

Clustering for Photo Retrieval at Image CLEF 2008

Diana Inkpen, Marc Stogaitis, François DeGuire and Muath Alzghool
School of Information Technology and Engineering
University of Ottawa
diana@site.uottawa.ca, mstog024@uottawa.ca, fdegu079@uottawa.ca, alzghool@site.uottawa.ca

This paper presents the first participation of the University of Ottawa group in the photo retrieval track at Image CLEF 2008. This year's task focused on clustering images in order to retrieve images from different clusters. We present our system, followed by results for the submitted runs. We worked only with the English part of the collection.

The research questions that we are investigating include: what happens if we index only the text captions, only the images, or the captions and the images; what is the performance of the system with and without clustering. We investigate different types of clustering. First, the k-means clustering algorithm, then hierarchical clustering in three variants: based on average link similarity, complete link, and single link. Then we try our own clustering method, based on searching words from the query and from the text caption in the WordNet[1] lexical knowledge base. We present four versions of this algorithm.

For text retrieval we used Lucene[2] and for image retrieval we use LIRE[3]. We have used a data fusion technique to merge the text retrieval with the image retrieval, usually giving more weight to the text retrieval results. We added a clustering component that clusters the text captions. Our text clustering component was implemented with the use of the Dragon[4] Toolkit.

Our WordNet-based clustering algorithm runs work as follows. The algorithm takes as input the ranked results list that was created by our standard text/image search. It then cycles through each word of the first document, looking for words that match the current clustering criteria (i.e., the words from the <cluster> field).

Our experiments show that text retrieval works well, and adding image similarity brings a bit of improvement. In terms of retrieving many different clusters, our WordNet-based algorithm worked best.

[1] http://wordnet.princeton.edu/
[2] http://lucene.apache.org/java/docs/
[3] http://www.semanticmetadata.net/lire/
[4] http://www.dragontoolkit.org/textcluster.asp

IPAL at CLEF 2008: Mixed-Modality-based Image Search, Novelty based Re-ranking and Extended Matching

Sheng Gao, Jean-Pierre Chevallet and Joo-Hwee Lim
IPAL, Institute for Infocomm Research, A*Star, Singapore
{gaosheng,viscjp,joohwee}@i2r.a-star.edu.sg

Our IPAL group has participated at CLEF 2008 on the new TEL collection and on the ad-hoc photographic retrieval ImageClef. Following the changes in evaluation criterion this year in ImageClef, i.e. promoting diversity in the top ranked images, we have integrated the novelty measure in our similarity based system developed in ImageCLEF 2007. The novelty score is calculated between an image in the ranked list and the images ranked higher than it. The system is still an automatic and mixed-modality based image search, which is similar to the previous years.

10 runs are submitted this year in ImageClef. In the overall ranking, our group stands at the 3rd place in 25 participants. However, the improvement using novelty measure is not significant when comparing with traditional similarity based system and the cluster identity of images cannot give us benefit as we expected.

4 runs are submitted for the TEL collection. For these runs and also 3 runs for ImageCLEF, we have used probabilistic links computed from Wikipedia. These links are used directly into the matching inner product. The results show no improvement using these links.

TIA-INAOE Participation at ImageCLEF 2008

H. Jair Escalante, J. A. González, Carlos A. Hernández, Aurelio López, Manuel Montes, Eduardo Morales, Luis E. Sucar and Luis Villaseñor
Research Group on Machine Learning for Image Processing and Information Retrieval
Department of Computational Sciences, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Luis Enrique Erro No. 1, 72840, Puebla, México
hugojair@ccc.inaoep.mx

This document summarizes the participation of INAOE's research group on machine learning for image processing and information retrieval (TIA) in the photographic retrieval task of ImageCLEF 2008. This year we proposed two approaches to the problem of image retrieval from multimedia collections.

First we studied annotation-based expansion of documents for image retrieval. This approach consists of automatically assigning labels to images by using supervised machine learning techniques. We used an in-development method for annotation that uses spatial relationships for improving the labeling. Labels were used for expanding the manual annotations of images. Then we built a text-based retrieval method that indexed the expanded annotations. Experimental results give evidence that the combination of labels and manual annotations could be helpful for improving retrieval performance and diversifying results of textual methods. However, it is not trivial to determine the best way for combining automatic annotations with the other information available.

In our second formulation we performed experiments with late fusion of heterogeneous methods. This approach consists of combining the outputs of independent retrieval methods of diverse nature and based on different sources. Our aim was to take advantage of the diversity, complementariness and redundancy of documents through ranked lists obtained with different methods and using distinct information. We considered content-based, text-based, annotation-based, visual-concept-based and multi-modal retrieval methods. For textual methods we used the full text in annotations and topics. For the visual-concept-based method we used the visual concepts of the XRCE group. The FIRE run was used as our content-based method. Intermedia-feedback, Web-based query expansion and annotation methods were used for the rest of the retrieval methods. A simple weighting scheme allowed us to effectively combine information from diverse sources. Despite the performance of independent retrieval methods is poor, the fusion of methods achieved competitive performance. Further, the heterogeneousness of the retrieval methods proved to be useful for diversifying the retrieval results. We report experiments with per-modality and hierarchical fusion. Better results were obtained with the latter strategy. For further diversifying the results of our methods we developed a simple strategy based on topic modeling with latent Dirichlet allocation. This technique resulted very helpful for some configurations, though it degraded the performance for others. This is mainly due to the quality of the initial retrieval results.

Effects of Visual Concept-based Post-retrieval Clustering in ImageCLEFphoto 2008

Masashi Inoue and Piyush Grover
National Institute of Informatics,Tokyo & Indian Institute of Technology, Kharagpur
m-inoue@nii.ac.jp, pgrover@cse.iitkgp.ernet.in

We examined the effectiveness of post-retrieval clustering that was based on the visual similarities among images to enhance the instance recall in the photo retrieval task of ImageCLEF 2008. The visual similarities are defined by the example visual concepts that were provided for the automatic photo indexing task. We tested two types of visual concepts and two kinds of clustering methods, hierarchical and modified k-means clustering. In all the runs, we used only the title fields in the search topics; we used either only the title fields or both the title and description fields of the annotations in English. The experimental results showed that hierarchical clustering can enhance instance recall while preserving the precision when certain parameters are appropriately set.

The Xtrieval Framework at CLEF 2008: ImageCLEF Photographic Retrieval Task

Thomas Wilhelm, Jens Kürsten and Maximilian Eibl
Chemnitz University of Technology Faculty of Computer Science, Dept. Computer Science and Media 09107
Chemnitz, Germany
[thomas.wilhelm | jens.kuersten | maximilian.eibl] at cs.tu-chemnitz.de

This paper describes our participation at the ImageCLEF photographic retrieval task. We used our Xtrieval framework for the preparation and execution of the experiments. This year, we submitted 4 experiments in total. The experiments showed that our thesaurus based query expansions works well in improving the geometric mean average precision (GMAP) and binary preference (BPREF), but deteriorates the improvements gained by the addition of content-based image retrieval. The baseline (text-only) scored a mean average precision (MAP) of 0.0998. The combination of text and image retrieval gained a raise by 37 percent to a MAP of 0.1364. After applying the query expansion to both experiments the MAP for the text-only retrieval increased to 0.1081, but the MAP for the combined text and image retrieval decreased to 0.1140. By implementing an interface to the PostgreSQL database the retrieval speed and comparison operations for vectors could be speeded up.

Different Multimodal Approaches Using IR-n in ImageCLEFphoto 2008

Sergio Navarro, Fernando Llopis and Rafael Muñoz
Natural Language Processing and Information Systems Group. University of Alicante, Spain. Spain
snavarro,llopis,rafael@dlsi.ua.es

This paper describes the approach of the University of Alicante to the problem of finding a suitable handling of multimodal sources within the ImageCLEF ad-hoc competition. We have worked on to add modifications to the most common multimodal techniques used in the image retrieval area in order to improve their performance. Moreover, we have added a clustering module in order to increase the number of different clusters that can be found within the top 20 images returned. Finally, we have studied the effect of using visual concepts in the retrieval phase and in the clustering phase. We can see in the results that with these multimodal techniques we have improved up to a 27% our results in a MAP way, respect the ones obtained using our last year configuration - a textual run using PRF -. Furthermore, we have seen that the use of LCA in a multimodal way outperforms clearly the MAP and P20 results obtained with other common methods used - it has obtained 0.3436 MAP, 4th place in the published task results, and 0.4564 P20, 5th place in the published task results -. Finally our TFIDF re-ranking run method has showed the best behaviour for the top 20 documents returned from our submissions, obtaining a F-measure value of 0.4051 - based on P20 and CR20 measures -. It makes us to conclude that the combination of these two mutimodal techniques will be the key for improving the performance in our system in future works.

Text-mess in the ImageCLEFphoto08 Task

S. Navarro[1], M.A. García[2], F. Llopis[1], M.C. Díaz[2], R. Muñoz[2], M.T. Martín[2], L.A. Ureña[2] and A. Montejo[2]
[1] Natural Language Processing and Information Systems Group. University of Alicante, Spain
{snavarro,llopis,rafael}@dlsi.ua.es
[2] Sistemas Inteligentes de Acceso a la Informaci´on, SINAI Group. University of Jaén, Spain
{magc,mcdiaz,maite,laurena,amontejo}@ujaen.es

This paper describes our participation in the ImagePhoto task at CLEF 2008. We present the joint work of two teams belonging to the TEXT-MESS project using a new system that combines the individual systems of these teams, one based on filtering and the other one based on clustering. We have submitted experiments using SINAI filtering method with the IR-n output, and the IR-n clustering module with the SINAI output. Our objective was to study the behaviour of these methods with a large number of configurations in order to increase our chances of success. The results show that a filtering method is not useful when we use the cluster terms or related words to filter retrieved documents, and that a clustering method can improve the results of cluster detection although at the expense of a decrease in precision of the results that is greater than the gain obtained for the CR20 measure with this method.

CLaC at ImageCLEFPhoto 2008

Osama El Demerdash, Leila Kosseim and Sabine Bergler
Computational Linguistics at Concordia (CLaC), Concordia university, Montreal.
{osama el,kosseim,bergler}@cse.concordia.ca

This paper presents our participation at the ImageCLEFPhoto 2008 task.

We submitted six runs, experimenting with our own block-based visual retrieval as well as with query expansion. For text retrieval, we used Apache Lucene search Engine.

The results we obtained show that despite the poor performance of the visual and text retrieval components, good results can be obtained through Pseudo-relevance feedback and the fusion of the results.

XRCE's Participation to ImageCLEF 2008

J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka and J.M. Renders
Xerox Research Centre Europe, 6 ch. de Maupertuis, 38240 Meylan, France
FirstName.LastName@xrce.xerox.com

This year, our participation to ImageCLEF 2008 (Photo Retrieval Sub-task) was motivated by trying to address three different problems: visual concept detection and its exploitation in a retrieval context, multimedia fusion methods for improved retrieval performance and diversity-based re-ranking methods. From a purely visual perspective, the representation based on Fisher vectors derived from a generative mixture model appeared to be efficient for both visual concept detection and content-based image retrieval. From a multimedia perspective, we used an intermediate fusion approach, based on cross-media relevance feedback that can be seen as a multigraph-based query regularization method with alternating steps. Finally, as one of main goals of the organizers was to promote both relevance and diversity in the retrieval outputs, we designed and assessed several re-ranking strategies that turned out to preserve standard retrieval performance (such at precision at 20 or mean average precision) while significantly decreasing the redundancy in the top documents. These re-ranking strategies were designed either as variant of the well-known maximal marginal relevance principle, or based on an explicit clustering algorithm.

The main lessons drawn from our participation to ImageCLEF-Photo were:

- in the case of pure text-based retrieval, both document and query enrichments by thesaurus improve the results, and combining the former with query expansion using pseudo-relevance feedback improves further the results;
- Fisher Vectors are rich image signatures and have state-of-the-art performance both in visual concept detection and content based image retrieval;
- the use of the visual concepts increases the retrieval performance when combined with pure text, but this advantage is lost when we use other, more complex multi-media fusion mechanisms, based on lower-level features than the visual concepts;
- combining the two mono-media information sources (image and text) using trans-media pseudo-relevance feedback improves significantly (by more than 50% relative) the retrieval results;
- concerning the diversity, most strategies that we proposed succeeded in reducing the redundancy in the top documents. As none of the techniques used explicitly the provided clustering criterion (e.g. diversifying according to cities or states or sports, etc.), the CR20 score was not always significantly increased (or in a few cases it was even decreased). This is not surprising, as we were seeking and improving the diversity in a blind (unsupervised) way.

Increasing Relevance and Diversity in Photo Retrieval by Result Fusion

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

This paper considers the strategies of query expansion, relevance feedback and result fusion to increase both relevance and diversity in photo retrieval.

In the text-based retrieval only experiments, the run with query expansion has better MAP and P20 than that without query expansion, and only has 0.85% decrease in CR20. Although relevance feedback run increases both MAP and P20, its CR20 decreases 10.18% compared with non-feedback run. It shows that relevance feedback brings in relevant but similar images, thus diversity may be decreased. The run with both query expansion and relevance feedback is the best in the four text-based runs.

In the content-based retrieval only experiments, the run without feedback outperforms the run with feedback. The latter has 10.84%, 9.13%, and 20.46% performance decrease in MAP, P20, and CR20.

In the fusion experiment, integrating text-based and content-based retrieval not only reports more relevant images, but also more diverse ones.

DCU at ImageCLEFPhoto 2008

Neil O'Hare, Peter Wilkins, Cathal Gurrin, Eamonn Newman, Gareth J.F. Jones and Alan F. Smeaton
Centre for Digital Video Processing, Dublin City University
nohare@computing.dcu.ie

DCU participated in the ImageCLEF 2008 photo retrieval task, submitting runs for both the English and Random language annotation conditions. Our approaches used text-based and image-based retrieval approaches to give baseline retrieval runs. The highest-ranked images from these baseline runs were clustered using K-Means clustering of the text annotations of the images. Finally, each cluster was represented by its most relevant image and these images were ranked for the final submission. Text retrieval used the BM-25 ranking algorithm with and without pseudo relevance feedback; image retrieval used a number of MPEG-7 features, which were combined using unsupervised query-time weight generation approaches. For text-based retrieval, we indexed the Title, Description, Notes and Location fields from the annotation documents, with the location field matched to a world gazetteer to automatically expand the location information to Town, State/County, Country and Continent.

For random language runs we using TextCat1 to identify German annotation documents, which were then translated into English using Systran Version:3.0 Machine Translator. In addition to translating the documents, we also submitted a number of runs which did not translate the annotation documents: although such runs are obviously handicapped in terms of retrieval, we were interested in exploring if these runs would achieve comparable performance to translated runs in terms of diversity.

Our results showed that, as expected, runs that combine image and text outperform text alone and image alone for general retrieval performance, and also for diversity. Our baseline image and text retrieval runs (i.e. without clustering) give our best MAP score, and these runs also outperformed the global mean and median of all ImageCLEFPhoto submissions for CR@20 and P@20. Clustering approaches gave a large improvement in CR@20 over the baseline, with an improvement of 22% for the monolingual text and image run, although P20 and MAP performance both suffer if clustering is used. Although pseudo relevance feedback consistently improved MAP, this improved retrieval came at the expense of diversity, as CR@20 was always lower when pseudo relevance feedback was used. We also found that untranslated random runs were able to achieve similar performance for diversity to translated random runs.

SZTAKI @ ImageCLEF 2008: Visual Concept Detection

Bálint Daróczy, Zsolt Fekete and Mátyás Brendel
Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{daroczyb, zsfekete, mbrendel}@ilab.sztaki.hu

We describe our approach to the ImageCLEF-VisualConcept 2008 task. Our method is based on image segmentation, using a feature vector describing the visual content of image segments or the entire image. The features include RGB average color and histogram, a 10x10 shape descriptor as well as a 2D Fourier transform of the image or segment. We used logistic regression for classification. Images were segmented by a home developed segmenter. While in this preliminary report classification by global image features performed best, preliminary results suggest the importance of segmentation for certain classes. We are planning to provide improved analysis in the near future.

Profil Entropic Visual Features for Visual Concept Detection in CLEF 2008 Campaign

Herve Glotin and Zhongqiu Zhao
Laboratoire des sciences de l'information et des systemes
UMR CNRS & Universite' Sud Toulon-Var France
glotin@univ-tln.fr, zhongqiuzhao@gmail.com

In this task we used only visual information to implement the VCDT task. We define and compare two simple projection operators: the harmonic and arithmetic means. We then propose a new kind of compact features based on the entropy of pixels projection. These features, called Profil Entropy Features (PEF), were added to usual color means and variances, and then were fed to SVM classifiers for the detection of 17 visual concepts on the IARPR images during the CLEF 2008 campaign. The simple arithmetic mean projection is at the 4th best rank at the official test over 53 runs of around 20 laboratories. We show that the harmonic projection gives complementary information, and that its simple early fusion with arithmetic PEF yields to the third best rank system. Xerox system is the best, certainly including SIFT features and large reference images database (see Xerox paper in this workshop). The usual perceptual color histograms features, of around 200 dimensions, that have been partly used by UPMC (see workshop note) seem similar or little less discriminatory than PEF. Moreover, it's fast to compute PEF, with around 10 images finished per second on usual pentium.

MMIS at ImageCLEF 2008: Experiments Combining Different Evidence Sources

Simon Overell[1], Ainhoa Llorente[2,3], Haiming Liu[2], Rui Hu[2], Adam Rae[2], Jianhan Zhu[2], Dawei Song[2]
and Stefan Rüger[2,1]

[1] Multimedia & Information Systems
Department of Computing, Imperial College London, SW7 2AZ, UK
[2] Knowledge Media Institute
The Open University, Milton Keynes, MK7 6AA, UK
[3] INFOTECH Unit
ROBOTIKER-TECNALIA, Parque Tecnológico, Edificio 202
E-48170 Zamudio, Bizkaia, Spain

seo01@doc.ic.ac.uk, jianhanzhu@gmail.com, {a.llorente, h.liu, r.hu, a.rae, d.song, s.rueger}@open.ac.uk

This paper presents the work of the MMIS group at ImageCLEF 2008. The results for three tasks are presented: Visual Concept Detection Task (VCDT), ImageCLEFphoto and ImageCLEFwiki. We combine image annotations, CBIR, textual relevance and a geographic filter using our generic data fusion method. We also compare methods for BRF and clustering.

Our top performing method in the VCDT enhances supervised learning by modifying probabilities based on a matrix that shows how terms appear together. Although it occurred in the top quartile of submitted runs, the enhancement did not provide a statistically significant improvement.

In the ImageCLEFphoto task we demonstrate that evidence from image retrieval can provide a contribution to retrieval; however we are yet to find a way of combining text and image evidence in a way to provide an improvement over the baseline.

Due to the relative performances of difference evidences in ImageCLEFwiki and our failure to improve over a baseline we conclude that text is the dominant feature in this collection.

ImageCLEFmed

CLEF2008 Image Annotation Task: an SVM Confidence-Based Approach

Tatiana Tommasi, Francesco Orabona and Barbara Caputo
Idiap Research Institute
Centre Du Parc, Rue Marconi 19, P. O. Box 592, CH-1920 Martigny, Switzerland
{ttommasi, forabona, bcaputo}@idiap.ch

This paper presents the algorithms and results of our participation to the medical image annotation task of ImageCLEFmed 2008. Our previous experience suggests that combining multiple cues with different SVM-based approaches is very effective in this domain. Moreover it points out that local features are the most discriminative cues for the problem at hand. So we decided to integrate two different local structural and textural descriptors: a modified version of the Scale Invariant Feature Transform (mod-SIFT) and Local Binary Pattern (LBP). They are combined through concatenation of the feature vectors and through the Multi-Cue Kernel.

The challenge this year consisted in annotating images coming mainly from classes with only few examples in the training set. We tackled the problem on two fronts. First, we used SVM as an opinion maker combining the first two opinions on the basis the confidence of the classifier's decisions. This approach produces class labels with "don't know" wildcards in it. As second strategy, we enrich the poorly populated training classes adding virtual examples generated modifying the original images.

Our team was called "idiap". We submitted 9 runs. Two of them consisted in repeating the 2007 winner run and applying on it the confidence based opinion fusion. We also submitted two separated runs for the modSIFT and the LBP features. The remaining runs consisted in:

- using cue-integration on the new features;
- combining cue-integration with the confidence based opinion fusion;
- combining cue-integration with the introduction of virtual examples in the training set;
- combining cue-integration with the confidence based opinion fusion and the introduction of virtual examples in the training set.

The results show that the classification performance increases passing from a single cue (idiap-LBP score 128.58; idiap-SIFTnew score 100.27) to multiple cues (LOW_lbp_siftnew score 93.20), from a hard decision (idiap-MCK pix sift score 313.01; LOW_lbp_siftnew score 93.20) to a soft decision through confidence based opinion fusion (idiap-MCK_pix_sift_2MARG score 227.82; LOW_2MARG score 83.79) and gets even better adding virtual examples in low populated classes (idiap-LOW_MULT_2MARG score 74.92). The run using jointly the low cue-integration technique, the confidence-based opinion fusion and the virtual examples ranked first among all submissions.

Automatic System for Extraction of Content-Based Characteristics from Digital Images

Gonzalo León, José Luis Delgado, Covadonga Rodrigo, Fernando López and Valentín Sama
Dpto. Lenguajes y Sistemas Informáticos
E.T.S.I. Informática - U.N.E.D, Spain.

In this paper we expose the development of a CBIR system (Content-based Image Retrieval) that is able to retrieve images from a corpus based upon the image content. In order to obtain such functionality, the system establishes a set of characteristics, which will be automatically generated. This allows the system to univocally identify each image from the collection. The sort of characteristics is diverse and they are related to concepts such as entropy, Gabor filters and image size. After the calculation of characteristics of each image, a calibration process is performed, whereby the system estimates the best weight for each characteristic. This estimation makes use of a calibration algorithm and a set of experiments, and the result is the influence of each characteristic in the main function that is used for the retrieval process. The calibration process starts in an equally balanced situation (all the characteristics have the same influence in the main function), and after several iterations the weight for each characteristic is fixed. The following task is the image validation, where the modifications to the main function are verified so as to ensure that the new function is better than the previous one. Finally, the image retrieval process is performed according to the ImageCLEFmed rules. The retrieval results have not been the expected ones, but we must say they are a good starting point that makes us establish several work lines for the future.

Multi-Relation Modeling on Multi Concept Extraction: LIG Participation at ImageClefMed

Loïc Maisonnasse, Eric Gaussier and Jean-Pierre Chevallet
Laboratory LIG
{loic.maisonnasse, eric.gaussier, jean-pierre.chevallet}@imag.fr

This paper presents the LIG contribution to the CLEF 2008 medical retrieval task (i.e. ImageCLEFmed). The main idea behind our contribution is to incorporate knowledge in the language modeling approach to information retrieval (IR). On ImageCLEFmed our model makes use of the textual part of the corpus and of the medical knowledge found in the Unified Medical Language System (UMLS) knowledge sources. Last year, we used UMLS to create a conceptual representation for each sentence in the corpus, and proposed a language modeling approach on these representations. The use of a conceptual representation allows the system to work at a more abstract semantic level, which solves some of the information retrieval problems, as the one of terminological variation. We also used different concept extraction methods, and tested how to combine these extraction methods on queries.

This year, we have extended our previous method in two ways: first, we have used, in addition to relations derived from UMLS, co-occurrence relations; second, we have combined concept extraction methods not only on queries, but also on documents. In this paper, we first detail some IR approaches that use advanced index terms. We then develop the graph model used in our submission to ImageCLEFmed 2008, and the different ways use to combine graphs derived from different concept extraction methods. After this, we present our results on this year collection, showing that combined concept extraction on document improves the MAP results and that relations impact more first results precision. Finally, we conclude this work and present some possible extensions.

TAU MIPLAB at ImageClef 2008

U. Avni[1], J. Goldberger[2] and H. Greenspan[1]
[1]Tel-Aviv University
[2] Bar-Ilan University

This paper describes the participation of Tel Aviv University Medical Image Processing Laboratory group at the ImageClef 2008 medical retrieval and medical annotation tasks. In both tasks we have used the bag-of-words approach for image representation. We submitted two purely visual automatic runs to the medical retrieval task, which used different normalization in the feature extraction stage. Images were converted to a histogram of visual words, and were compared using L1 distance. Our best run was ranked first among the automatic visual based retrieval systems, with MAP score of 0.042. For the medical annotation task we submitted four runs, all used support-vector-machines trained on the visual word histograms. The runs differ in image resolution, and in the way classifiers of two resolutions were combined. In this task our result was second best among the participating groups, with error scores between 105.75 and 117.17.

MedGIFT at ImageCLEF 2008

Xin Zhou[1], Julien Gobeill[1] and Henning Müller[1 2]
[1] Medical Informatics, Geneva University Hospitals and University of Geneva, Switzerland
[2] University of Applied Sciences Western Switzerland (HES SO), Sierre, Switzerland
xin.zhou@sim.hcuge.ch

This article describes the participation of the Geneva University Hospitals and the University of Geneva at the 2008 ImageCLEF image retrieval benchmark. We concentrated on the two tasks concerning medical imaging: image retrieval from medical collections and medical image annotation. The visual information analysis is based on the GNU Image Finding Tool (GIFT). Other information such as textual information and aspect ratio are integrated to improve the results. The main techniques are the same as in past years, with a little tuning to slightly improve results.

For the image retrieval task, 3 purely visual runs and 5 mixed-media automatic runs were submitted. One of the purely visual runs (GIFT4) used the same technique during the past five years to provide a baseline. Best results among the purely visual runs is (GIFT8), with a MAP of 0.0349 and a precision at 10 of 0.17. One textual run (HUG{BL{EN) was provided by another research group of the Unievrsity hospitals. Various strategies for a combination of visual and textual runs were tested (GIFT8) and (HUG{BL{EN) to improve the results. The best MAP is obtained by simply combining textual and visual runs with equal weight (GIFT8 EN0.5) resulting in a MAP of 0.0848. Compared to the original text runs, the combination with our visual run improves early precision slightly, but reduces MAP significantly.

For the medical image annotation task, the basic GIFT system was used for the feature extraction. The work of this year followed work performed in 2007. In 2008, we added two other factors: the frequency of images of each class in the training data and the hierarchy information inside of each axis of the IRMA code. Submitted runs used either kNN approach or a voting{based approach for classification. A dynamic kNN approach took into account the frequency of images of each class and obtained the best performance among the runs based on the kNN approach. Classifying the code per axis using a voting{based approach gives the best overall result. Further investigation proves that using the hierarchy information inside each axis and classifying the axis recursively can improve the results if a high threshold value is applied. The best result obtained an error score of 181.17.

For the visual tasks it becomes clear that the baseline GIFT runs do not have the same performance as more sophisticated modern techniques do. Due to time constraints no optimizations could be performed and no relevance feedback was used, usually one of the strong points of GIFT.

MIRACLE at ImageCLEFmed 2008: Evaluating Strategies for Automatic Topic Expansion

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3] and José C. González-Cristóbal[1,3]
[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

This paper describes the participation of MIRACLE research consortium at the ImageCLEFmed task of ImageCLEF 2008. The main goal of our participation this year was to compare among different topic expansion approaches: methods based on linguistic information such as thesauri or knowledge bases, and statistical techniques mainly based on term frequency. Thus we focused on runs using text features only. All experiments were fully automatic, with no manual intervention.

The architecture of our system is composed of four different modules: the textual (text-based) retrieval module, which indexes medical case descriptions in order to search and find the most relevant ones to the text of the topic; the expander module, which performs the expansion of the content of documents and/or topics with related terms using textual and/or statistical methods; the relevance-feedback module, which allows to execute reformulated queries that include the results of a initial seed query; and, finally, the result combination module, which uses OR operator to combine, if necessary, the results provided by the previous subsystems.

Instead of using raw terms, the textual information of both documents and topics is parsed and tagged to unify all terms into concepts of medical entities. The result is that medical concept identifiers 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 meta-thesaurus.

For all experiments, a common baseline algorithm was used to process the document collection. This algorithm is based on the following steps that are executed sequentially: text extraction, medical-vocabulary recognition, tokenization, conversion to lowercase, filtering, stemming and, finally, indexing and retrieval.

This common baseline algorithm is then complemented and combined with different expansion techniques in order to compare the improvement given by semantic- vs. statistical-based techniques. For the semantic expansion, we used the MeSH concept hierarchy using the UMLS entities detected in document and topics as basic root elements to expand with their hyponyms (i.e., other entities whose semantic range is included within that of the root entity). The statistical method consists of using Agrawal's apriori algorithm to expand the topics with the terms in the consequents of discovered rules, i.e., UMLS entities that are related to the topic according to the document corpus. In addition, relevance-feedback techniques were also used.

We submitted 8 runs. The highest MAP (0.266) is obtained with the baseline experiment in English, which is on the average with respect to other participating groups for purely textual experiments. Moreover, MAP values are similar in practice for experiments using topic expansion, and noticeably worse (0.105 vs. 0.266) in the case of relevance-feedback. This shows that no strategy for either topic expansion or specially relevance-feedback has proved to be useful. The re-ranking algorithm used for combining the different result lists is likely to be the main reason for the disappointing results. Another probable cause is the choice of the OR operator to combine the terms in the topic to build up the query. We think that MAP values might be significantly higher using this operator.

MIRACLE at ImageCLEFannot 2008: Classification of Image Features for Medical Image Annotation

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3], José Carlos González-Cristóbal[1,3] and José Miguel Goñi-Menoyo[1]

[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

This paper describes the participation of MIRACLE research consortium at the ImageCLEF Medical Image Annotation task of ImageCLEF 2008. While in previous participations we approached this task as a machine learning problem, regardless of the domain, as our areas of expertise did not include image analysis research, a lot of effort was invested last year to develop our own image analysis system, based on MATLAB, to be used in our experiments. Thus, now the focus of our experiments is to test and evaluate this system in-depth and make a comparison among diverse configuration parameters such as number of images for the relevance feedback to use in the classification module.

Our system has two functional blocks. The first block is the feature extraction module, which is in charge of the calculation and extraction of image features. This module has been entirely developed using MATLAB and extracts vectors with a total of 3,741 features for each image. Images are first converted to gray-scale and rescaled to 256x256 pixels. Then the system extracts a variety of global and local features including gray histogram (128 levels of gray), image statistics (mean, median, variance, maximum singular value, skewness and kurtosis), Gabor features (4 scales, 6 filter orientations), fractal dimension, Discrete Cosine Transform (DCT) coefficients, Discrete Wavelet Transform (DWT) coefficients, Tamura features (coarseness, contrast, directionality), and co-occurrence matrix statistics (energy, entropy, contrast, homogeneity, correlation). For local features, images are cut up into 64x64 pixel blocks and then the same features are extracted for each block.

The second block is the classifier, which determines the IRMA code associated to a given image, from its feature vector and the feature matrix of the training set. The classifier is internally composed of two blocks: one for selecting those images in the training set whose vectors are at a distance lower than a given threshold from the vector associated to the image to classify, and another that actually generates the IRMA code, depending on the codes and similarity of nearby images.

We submitted four runs. For all of them, the returned IRMA code was generated from the combination of the first N images in the training set that are most similar to the image to classify. The combination consists of a simple "addition" of strings characters in which, if both characters are different, the result is the wildcard "*" representing the ambiguity. This algorithm actually could be considered as a variation of the classical k-Nearest Neighbour algorithm with a specific definition of the generating the output class. Additionally, two runs used relevance feedback (RF).

The best score is achieved by combining the codes of the first 3 images, with no relevance feedback. Comparing to other participants in the task, we achieve average results and rank 4th out of 6 groups.

Due to a mistake when carrying out the experiments, the calculation of the distance among vectors assigns the same weight to every vector dimension, regardless of the nature of the feature to which this component belongs and/or the number of components belonging to that feature. Obviously, the feature matrix should have been divided into different feature sub-matrixes and we should have employed different distances for calculating similarities and combined them using different weight strategies. Of course, this will be solved for future participations.

Text-mess in the Medical Retrieval ImageCLEF08 Task

S. Navarro[1], M.C. Díaz[2], R. Muñoz[1], M.A. García[2], F. Llopis[1], M.T. Martín[2], L.A. Ureña[2] and A. Montejo[2]

[1]Natural Language Processing and Information Systems Group. University of Alicante, Spain

{snavarro,rafael,llopis}@dlsi.ua.es

[2]Sistemas Inteligentes de Acceso a la Informaci´on, SINAI Group. University of Jaén, Spain

{mcdiaz,magc,maite,laurena,amontejo}@ujaen.es

This paper describes our participation in the Medical Retrieval task at ImageCLEF 2008. We present the joint work of two teams belonging to the TEXT-MESS project using a new system that combines the 2 individual systems of these teams. The aim of the experiments performed is to figure out if there are techniques used in one of the two systems which can complement the other system in order to improve their performance. The best results obtained in the training phase and in the competition have been reached with a configuration which uses the IR-n system with a negative query expansion based on the acquisition type of the image mixed with the SINAI system with a MeSH based query expansion. We have obtained a MAP of 0.2777 for our best run, obtaining the 5th place in the ranking of textual participant runs submitted, and the 6th place in the global classification.

SINAI at ImageCLEFmed 2008

M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia, L.A. Ureña-López and A. Montejo-Ráez

University of Jaén. Computer Science Department
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain

{mcdiaz,magc,maite,laurena,amontejo}@ujaen.es

This abstract describes the SINAI team participation in the ImageCLEF campaign. In this abstract we only explain the experiments accomplished in the medical task. We have experimented with query expansion and the text information of the collection. For expansion, we carry out experiments using MeSH ontology and UMLS separately. To expand the queries with UMLS we have used MetaMap program. With respect to text collection, this year, a new collection have been used. It contains images from articles published in Radiology and Radiographics including the text of the captions and a link to the HTML of the full text articles. We have downloaded articles from the web and constructed a new textual collection including the text of the article section where the image appears. We have used three different collections, one with caption and title, other with caption, title and the text of the section where the image appears, and the third with the full article. Moreover, we have experimented with mixed search, textual and visual search, using the FIRE software for image retrieval.

Our main goal is to investigate the effectiveness of different expansions and different sizes in textual collections. Moreover, we have experimented with the influence of mixing visual information with our results. The visual results have been obtained with the FIRE software. To mix textual and visual results, we have used the same algorithm that applied in 2007. In previous years the best results reached were obtained with a weight of 0.8 for textual results and 0.2 for visual ones. This year we have only experimented with these weights.

The use of FIRE and MeSH expansion with the minimal collection (only caption and title) obtains the best results in the track. The use of UMLS expansion obtains worse results than the baseline.

A Multimodal Approach to the Medical Retrieval Task Using IR-n

Sergio Navarro, Rafael Muñoz and Fernando Llopis
Natural Language Processing and Information Systems Group. University of Alicante, Spain.
{snavarro,rafael,llopis}@dlsi.ua.es

In our participation in the Medical Retrieval task we wanted to figure out if good results can be achieved with IR-n - our IR passage based system - for this restricted domain. We have focused on comparing the behaviour of two relevance feedback methods in this task - LCA and PRF -. Furthermore, in order to adapt our system to this task we have used two automatic query expansion techniques related with the medical domain. On one hand we have added to our system an automatic query expansion method based on MeSH ontology and on the other hand we have added a negative query expansion based on the acquisition type of the image. Finally we have added a multimodal re-ranking module - late fusion -. We have used two operation modes, one merges the two list in a classical re-ranking way, and the other mode bases the calculus of the relevance of an image on the quantity and the quality of the text related to the image in order to take the decision as to which system is more confident for that image - the system based on text or the one based on images -. A major finding of the results is that our passage based system fits very well to this task. Within the textual runs submitted by all the participants we have reached the 6th place for our baseline and the 1st place for a run using PRF and query expansion adapted to the medical domain. Our results for multimodal re-ranking have not been successful due to problems with the parameters tuning for the test collection of this year.

Text-Only Cross-Language Image Search at Medical ImageCLEF 2008

Julien Gobeill, Patrick Ruch and Xin Zhou
University and Hospitals of Geneva, Switzerland
julien.gobeill@sim.hcuge.ch

We report on simple textual strategies with thesaural resources in order to perform document and query translation for cross-language information retrieval in a collection of annotated medical images. The keystone of our strategy for the previous medical ImageCLEF was to enrich documents and queries with Medical Subject Headings (MeSH) terms extracted from them, in order to translate the more important concepts into an intermediate language. The core technical component of our cross-language search engine is an automatic text categorizer, which associates a set of MeSH terms to any input text, with a top precision at above 90%. Nevertheless, in the new 2008 collection, images are given with more verbose captions, and with an associated article relative to a specific case study. Therefore, our strategy to enrich each document is either to collect MeSH terms from the associated article, either to extract them from the caption. Our results are fair, as we stand on the first part of the participants (0.176 for mean average precision). Nevertheless, it appears that MeSH terms collected from the relative article are not always relevant, as this article can concern a huge set of images in general, and can not to describe precisely the associated image. Moreover, the MeSH terms directly extracted from the captions lead to worst performances, possibly due to the more verbose captions. We try different strategies on weighting scheme or retrieval on articles, but without significant improvements. In conclusion, a mixed strategy to combine the two origins of the MeSH terms should be planned for the next ImageCLEF, while better performances should be obtained in the future by tuning the system with the existing benchmark.

Multimodal Medical Image Retrieval: OHSU at ImageCLEF 2008

Jayashree Kalpathy-Cramer, Steven Bedrick, William Hatt and William Hersh
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University, Portland, OR, USA
{kalpathy,bedricks,hattb,hersh}@ohsu.edu

We present results from Oregon Health & Science University's participation in the medical image retrieval task of ImageCLEF 2008. We created a web-based retrieval system built on a full-text index of the annotations using a Ruby on Rails framework. The text-based search engine was implemented in Ruby using Ferret, a port of Lucene. 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. Our system provides the user with a number of search options including those for limiting the search to the desired modality, UMLS-based term expansion and Natural Language Processing based techniques. Purely textual runs as well as mixed runs using the purported modality were submitted. We also submitted interactive runs using a number of user specified search options. Latent semantic analysis of the visual features was used to reorder results.

The use of the UMLS Metathesaurus increased our recall. However, our system is primarily geared towards precision. Consequently, many of our multimodal automatic runs using the custom parser as well as interactive runs had high early precision. Our runs also performed well using the Bpref metric, a measure that is more robust in the case of incomplete judgments. Our best mean average precision was 0.23, our best precision at 10 was 0.55 (the highest overall) and our best Bpref was 0.35, the second highest overall.

Baseline Results for the CLEF 2008 Medical Automatic Annotation Task

Mark O. Gueld and Thomas M. Deserno
Department of Medical Informatics, RWTH Aachen University, Aachen, Germany

This work reports baseline results for the CLEF 2008 Medical Automatic Annotation Task (MAAT) by applying a classifier with a fixed parameter set to all tasks 2005-2008. The classifier performs a weighted combination of three distance and similarity measures operating on global image features: Scaled-down representations of the images are compared via metrics that model the typical variability in the image data, mainly translation, local deformation, and radiation dose.

In addition, a distance measure based on texture features is used. For classification, a k nearest neighbor classifier is used. In 2008, the baseline classifier yields error scores of 170.34 and 182.77 for k=1 and k=5 when the full code is reported, which corresponds to error rates of 51.3% and 52.8% for 1-NN and 5-NN, respectively.

Judging the relative increases of the number of classes and the error rates over the years, MAAT 2008 is estimated to be the most difficult in the four years.

Methods for Combining Content-Based and Textual-Based Approaches in Medical Image Retrieval

Mouna Torjmen, Karen Pinel-Sauvagnat and Mohand Boughanem
SIG-IRIT-Toulouse-France
{Mouna.Torjmen , Karen.Sauvagnat , Mohand.Boughanem }@irit.fr

Our aim of participating in the Medical Image Retrieval task of Image CLEF 2008 was to evaluate different combination methods for purely textual and visual approaches. Indeed, we compare a classical combination method using a linear combination function to a combination method taking into account the query type: visual, textual, mixed. The used systems for content-based image retrieval and textual-based image retrieval are respectively GIFT and XFIRM.

As the document structure in the collection is not complex, we use a simplified version of the XFRIM model. This latter is based on a relevance propagation method. During query processing, relevance scores are computed at leaf nodes level and then at inner nodes level thanks to a propagation of leaf nodes scores through the document tree. An ordered list of sub-trees is then returned to the user. Only two fields of the document in the medical textual collection are indexed (“caption” and “title”) as they are the only clues (elements), which contain significant textual information.

In our experiments, we directly used GIFT results kindly provided by organizers, with no further processing.

Classical combination. We used the two aforementioned systems on the whole set of queries. To merge the result lists into a single list of ranked results, we first normalize scores obtained by the two systems, and then use a simple and classic linear combination function as follows:

- Combination according to query type: we evaluated the processing of each query category with a different system. We thus used the GIFT system to evaluate visual topics, the XFIRM system to evaluate textual topics, and a classic combination function of the two systems to evaluate mixed topics.

Results. Our most interesting result is that combining scores provided by both systems using classical combination function allows to obtain higher retrieval accuracy in terms of MAP measure than the combination according to the query type (MAP=0.1705 with α=0.9 versus MAP=0.1101). Moreover, it is more reliable than using only textual retrieval (MAP=0.1410) or using only visual retrieval (MAP=0.0349). So, visual information could be used as an additional source of evidence to improve results but not as a solely information source.

ImageCLEFWiki

UPMC/LIP6 at ImageCLEF's WikipediaMM: An Image-Annotation Model for an Image Search-Engine

Ali Fakeri-Tabrizi, Massih-Reza Amini, Sabrina Tollari and Patrick Gallinari
Université Pierre et Marie Curie-Paris 6, Laboratoire d'Informatique de Paris 6 - UMR CNRS 7606, 104 avenue du président Kennedy, 75016 Paris, France
firstname.lastname@lip6.fr

In this paper, we present the LIP6 retrieval system which automatically ranks the most similar images to a given query constituted of both textual and/or visual information through a given textual-visual collection. The system first preprocesses the data set in order to remove stop-words as well as non-informative terms. For each given query, it then finds a ranked list of its most similar images using only their textual information. Visual features are then used to obtain a second ranking list from a manifold and a linear combination of these two ranking lists gives the final ranking of images.

CWI at ImageCLEF 2008

Theodora Tsikrika, Henning Rode and Arjen P. de Vries
CWI, Amsterdam, The Netherlands
{Theodora.Tsikrika, Henning.Rode, Arjen.de.Vries}@cwi.nl

CWI participated in the wikipediaMM task at ImageCLEF 2008. Our aim was to examine the value of textual evidence for the retrieval of Wikipedia images associated with sparse and noisy English text formatted in XML and establish a strong text-based baseline against which to compare results of future experiments.

To this end, we employed a text-based language modelling approach by considering that the text associated with each image corresponds to a textual document and that queries consist only of the topics' textual part (the title field). The multinomial language model we used for retrieval is based on query likelihood, with the individual term probabilities estimated using maximum likelihood estimates and smoothed using a mixture model of the document model with a background model (the collection model in this case). Ranking was produced by the posterior probability of a document being relevant to a query, so that prior probabilities given the documents' query-independent features can be incorporated.

We used PF/Tijah (http://dbappl.cs.utwente.nl/pftijah/), a flexible open source XML retrieval system for indexing (with stopword removal and stemming) and retrieval. We submitted two runs based on the smoothed language model, one with a uniform prior and one that incorporates a prior based on a linear function of length, so as to bias retrieval towards images with richer descriptions. Length was defined as the number of terms in the image description. Both these text-based runs performed satisfactorily, with the former run performing slightly better than the latter. We performed a retrospective analysis of the distribution of length in the wikipediaMM collection and the relevant images for the 2008 topics: the collection mostly contains images with shorter descriptions, while the relevant ones appear to be associated with slightly longer descriptions. However, the smoothed language modelling approach with uniform prior already retrieves documents with sizes similarly distributed to those of the relevant ones. Thus further biasing towards images with richer descriptions is not beneficial.

Some Experiments on the WikipediaMM 2008 Task: Evaluating the Impact of Image Names in Context-Based Retrieval

Mouna Torjmen, Karen Pinel-Sauvagnat and Mohand Boughanem
IRIT-Paul Sabatier University- Toulouse-France
{Mouna.Torjmen, Karen.Sauvagnat, Mohand.Boughanem}@irit.fr

The goal of our participation in the WikipediaMM task of CLEF 2008 was to study the use of the name of images in a context-based retrieval approach. Our intuition behind this study is that the name of an image, if it is significant, describes well the image content, and consequently, it plays an important role in determining the image relevance.

The XFIRM model. As the document structure in the collection is not complex, we use a simplified version of the XFIRM model. This latter is based on a relevance propagation method. During query processing, relevance scores are computed at leaf nodes level and then at inner nodes level thanks to a propagation of leaf nodes scores through the document tree. An ordered list of sub-trees is then returned to the user.

Algorithms used to evaluate the impact of image names

Using only the image name terms. To explicitly use the terms composing the image name, we propose to only use the image name keywords to retrieve relevant images. We compute a score for each image using the vector space model. We evaluated 3 similarity measures: the Cosine Similarity, the Dice Coefficient and the Inner Product.

Combining image name scores and document scores. The score of the document (image) obtained using the image name W_{ImName}(doc) could be combined with the score obtained by the XFIRM system λW_{XFIRM}(doc).

$$W(doc) = \lambda W_{XFIRM}(doc) + (1 - \lambda) W_{ImName}(doc)$$

Implicit use of the image name keywords. We modify the term weighting formula used in the XFIRM model, by increasing the score of terms in the image name, by means of multiplying the score with a factor K.

Results. For runs evaluated only using the image name, Cosine Similarity and Inner measures give approximately the same results. These latter are better than those using the Dice coefficient.

As a main conclusion, we notice that the image name is a relevant contextual element of image retrieval whatever the way it is introduced (implicitly: MAP =0.1724 with K=1.1, or explicitly: MAP=0.1681 with λ=0.9) comparatively to the use of each source of evidence separately (MAP= 0.0743 of image name only versus MAP= 0.1652 of textual information only).

Moreover, we conclude that the implicit use of image names is more interesting than the implicit use of this contextual element in contextual image retrieval.

The Xtrieval Framework at CLEF 2008: ImageCLEF Wikipedia MM task

Thomas Wilhelm, Jens Kürsten and Maximilian Eibl
Chemnitz University of Technology Faculty of Computer Science, Dept. Computer Science and Media 09107
Chemnitz, Germany
[thomas.wilhelm | jens.kuersten | maximilian.eibl] at cs.tu-chemnitz.de

This paper describes our participation at the ImageCLEF Wikipedia MM task. We used our Xtrieval framework for the preparation and execution of the experiments. We submitted 4 experiments in total. The results of these experiments were mixed. The text-only experiment scored second best with a mean average precision (MAP) of 0.2166. In combination with image based features the MAP dropped to 0.2138. With the addition of our thesaurus based query expansion it scored best with a MAP of 0.2195. Without query expansion and with the inclusion of the provided concepts the lowest MAP of 0.2048 was achieved, but there were 23 more relevant documents retrieved than in all 3 other experiments. Furthermore, the retrieval speed and comparison operations for vectors could be speeded up by implementing an interface to the PostgreSQL database.

A Textual Approach Based on Passages Using IR-n in WikipediaMM Task 2008

Sergio Navarro, Rafael Muñoz and Fernando Llopis
Natural Language Processing and Information Systems Group. University of Alicante, Spain.
snavarro,rafael,llopis@dlsi.ua.es

In this paper we have focused our efforts on comparing the behaviour of two relevance feedback methods in this task - LCA and PRF - and in checking if our passage based information retrieval (IR) system is useful in a competition with small sized documents. Furthermore we have added an adaptation to this domain based on decompound in single terms those file names which use a Camel Case notation. We base our decision on the belief that the most meaningful information of an image file appointed by a human is on the file name itself. Thus, it is important to make visible this terms when they are hidden in a compounded file name. Finally we have added a geographical query expansion and a visual concept expansion. We have obtained a 29th place within a total of 77 runs with our baseline run - which only used the passage IR system -, and a 3rd place obtained with our best run - which used the passage IR system with Camel Case decompounding -. It shows us on one hand the usefulness of our passage based IR system in this domain, and on the other hand it confirms our belief in the existence of especially meaningful information within the file names. In the relevance feedback respect, we have obtained contradictory results about the suitability of LCA or PRF to the task, but we have found that LCA has a more robust behavior than PRF.

Increasing Cluster Recall of Cross-Modal Image Retrieval

Simon Rácz, Bálint Daróczy, Dávid Siklósi, Attila Pereszlényi, Mátyás Brendel and András Benczúr
Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{sracz, daroczyb, sdavid, peresz, mbrendel, benczur}@ilab.sztaki.hu

We describe our approach to the ImageCLEF Photo and WikiMediaMM 2008 tasks. The novelty of our method consists of combining image segment based image retrieval with our text based approach. We rank text hits by our own Okapi BM25 based information retrieval system and image similarities by using a feature vector describing the visual content of image segments. Images were segmented by a home developed segmenter. We use automatic query expansion by adding new terms from the top ranked documents. Queries were generated automatically from the title and the downweighted description words.

Conceptual Image Retrieval over the Wikipedia Corpus

Adrian Popescu, Hervé Le Borgne and Pierre-Alain Moëllic
CEA-LIST
LIC2M (Multilingual Multimedia Knowledge Engineering Laboratory)
B.P.6 – F92265 Fontenay-aux-roses Cedex, France
{adrian.popescu, herve.le-borgne, pierre-alain.moellic}@cea.fr

Image retrieval in large-scale databases is currently based on a textual chains matching procedure, a technique that produces good results as long as the annotations associated to pictures are accurate and detailed enough. These conditions are not met for a large majority of image corpuses, such as the Wikipedia collection, and it is interesting to explore methods that go beyond chain matching. In this paper, we present our approach to image retrieval, tested in the ImageCLEF 2008 WikipediaMM. The approach is based on a query reformulation using concepts that are semantically related to those in the initial query. For each interesting entity in the query, we used Wikipedia and WordNet to extract and list of related concepts, which were further ranked in order to propose the most salient in priority. We also made a list of visual concepts which were used in order to re-rank the answers to queries that included, implicitly or explicitly, these visual concepts. The CEA submitted two automatic runs, one based on query reformulation only and one combining query reformulation and visual concepts, which were ranked 4th and 2nd using the MAP measure.

Concept Content Based Wikipedia WEB Image Retrieval Using CLEF VCDT 2008

Zhongqiu Zhao and Herve Glotin
LSIS - Laboratoire des sciences de l'information et des systemes
UMR CNRS & Universite' Sud Toulon-Var France
glotin@univ-tln.fr, zhongqiuzhao@gmail.com

One challenge for this Wikipedia task is the training of visual models. We propose in this paper to link each topics one or few visual concepts of the Visual Concept Detection (VCDT) CLEFimage08 task, even if three topics do not fit VCDT concepts. We use the same models and features than in our VCDT systems. We show that our visual IMG NOFB run is the second best model in this campaign for this run type. So it can be concluded that our VCDT visual concept partly fit this task. Moreover we show that even a simple boolean text analysis overcomes the best IMG NO FEEDBACK run, which has 0.0037 MAP, against 0.399 for our TXT NOFB text run. This emphasizes the fact that visual retrieval for Wiki task is very difficult. This can explain why our 3 basic fusions methods TXTIMG did not improve the TXT run. One can then conclude that the Feedback seems necessary for such task.

PKU at ImageCLEF 2008: Experiments with Query Extension Techniques for Text-Base and Content-Based Image Retrieval

Zhi Zhou, Yonghong Tian, Yuanning Li, Ting Liu, Tiejun Huang and Wen Gao
The Institute of Digital Media, School of EE & CS
Peking University, Beijing 100871, China
{zzhou, ynli, tliu}@jdl.ac.cn, {yhtian,tjhuang,wgao}@pku.edu.cn

In this paper, we present our solution for the WikipediaMM task at ImageCLEF 2008. The aim of WikipediaMM 2008 task is to investigate effective retrieval approaches in the context of a large-scale and heterogeneous collection of Wikipedia images that are searched by textual queries (and/or sample images and/or concepts) describing a user's information need. We first experimented with a text-based image retrieval approach with query extension, where the expansion terms are automatically selected from a knowledge base that is (semi-)automatically constructed from Wikipedia. We show how this open, constantly evolving encyclopedia can yield inexpensive knowledge structures that are specifically tailored to effectively enhance the semantics of queries. Encouragingly, the experimental results rank in the first place among all submitted runs. The second approach we experimented with is content-based image retrieval (CBIR), in which we first train 1-vs-all classifiers for all query concepts by using the training images obtained by Yahoo! search, and then model the retrieval task as visual concept detection in the given Wikipedia image set. By comparison, this approach performs better than other submitted CBIR runs. Finally, we experimented with a cross-media image retrieval approach by combining and re-ranking text-based and content-based retrieval results. Despite the final experimental results were not formally submitted before the submission deadline, this approach performs remarkably better than the single text-based or visual-based retrieval approaches.

UJM at ImageCLEFwiki 2008

Christophe Moulin, Cécile Barat, Mathias Géry, Christophe Ducottet and Christine Largeron
Université de Lyon
UMR 5516, Saint-Etienne, France
{Christophe.Moulin, Cecile.Barat, Mathias.Gery, Christophe.Ducottet, Christine.Largeron}@univ-st-etienne.fr

This paper reports our multimedia information retrieval experiments carried out for the ImageCLEF track (ImageCLEFwiki). The purpose of our experiments is twofold: firstly, our overall aim is to develop a multimedia document model combining text and/or image modalities. Secondly, we aim to compare results of our model using a multimedia query with a text only model.

In this paper, we introduce our model and we briefly describe indexing and retrieval processes. We present our initial results which demonstrate that visual information is useful as it allows to find documents that were not found with methods based on text only.

Our multimedia document model is based on a vector of textual and visual terms. The textual terms correspond to words. The visual ones result from local colour descriptors which are automatically extracted and quantized by k-means, leading to an image vocabulary. They represent the colour property of an image region. To perform a query, we compute a similarity score between each document vector (textual + visual terms) and the query using the Okapi method based on the tf.idf approach.

We have submitted 6 runs either automatic or manual, using textual, visual or both information. Thanks to these 6 runs, we aim to study several aspects of our model, as the choice of the visual words and local features, the way of combining textual and visual words for a query and the performance improvements obtained when adding visual information to a pure textual model. Concerning the choice of the visual words, results show us that they are significant in some cases where the visualness of the query is meaningful. The conclusion about the combination of textual and visual words is surprising. We obtain worth results when we add directly the text to the visual words. Finally, results also inform that visual information bring complementary relevant documents that were not found with the text query. These initial results are promising and encourage the development of our multimedia model.

Multilingual Web Track (WebCLEF)

Overview of WebCLEF 2008

Valentin Jijkoun and Maarten de Rijke
ISLA, University of Amsterdam
jijkoun,mdr@science.uva.nl

We describe the WebCLEF 2008 task. Similarly to the 2007 edition of WebCLEF, the 2008 edition implements a multilingual ``information synthesis" task, where, for a given topic, participating systems have to extract important snippets from web pages. We detail the task and the assessment procedure. At the time of writing evaluation results are not available yet.

On the Evaluation of Snippet Selection for Information Retrieval

A. Overwijk, D. Nguyen, C. Hauff, R.B. Trieschnigg, D. Hiemstra and F.M.G. de Jong
Twente University
arnold.overwijk@gmail.com, dong.p.ng@gmail.com, c.hauff@ewi.utwente.nl, trieschn@ewi.utwente.nl,
hiemstra@cs.utwente.nl, f.m.g.dejong@ewi.utwente.nl

In this paper we take a critical look at the evaluation method of WebCLEF 2007. The suitability of the evaluation method can be seen from two sides, namely from a participating system and a non participating system. A participant has the advantage that the evaluation is partly based upon his output. In this paper we will investigate if the size of the pool of snippets, the implementation of the evaluation method and the quality of the assessments is sufficient enough for reliable evaluation. We exploit bug in last year's best performing system to show that the pool of snippets is not large enough. To prove that the implementation of the evaluation measures (i.e. precision and recall) are not correctly implemented, we show that an output that is almost similar to last year's best performing system has a huge decrease in performance. In addition we show that the quality of the assessments is not suitable. Based on these results we have to conclude that the evaluation is inappropriate. Therefore some alternative evaluation methods will be discussed concluding in a recommendation to improve the evaluation of WebCLEF in the future.

REINA at WebCLEF 2008

Carlos G. Figuerola, José L. Alonso Berrocal, Angel F. Zazo Rodríguez and Montserrat Mateos
REINA Research Group, University of Salamanca, Spain
http://reina.usal.es, reina@usal.es

This year, the WebCLEF track is similar to the 2007 edition, that is: retrieving text snippets or fragments of web pages which bring up information about a topic; additionally, snippets must be in a language from a set of accepted ones. As in 2007, we have a set of topics, each with a title and a short description, as well as several documents or 'known sources' about the topic. Additionally, for each topic, we have one or several searches in Google, with the first 1000 documents retrieved.

In our approach, for each topic we considered all documents retrieved after queries to Google as the collection of documents with which to work. These documents are to be fragmented into pieces, each of whom will be treated as a separate document. For the queries, we use the description that we have for each topic. This query can be enriched with more terms from the 'known sources'. So, the task can be approached like a classic problem of retrieval, and apply, consequently, conventional techniques.

To segment documents, as we wanted fragments that had informative sense, our fragmenter looks for the period closest the 1500 bytes, and part by that point. However, web pages are not conventional documents; many web pages are viewed by the user as a set of visual blocks that have different functions and containing different types of information. The conventional tools of conversion to plain text are not able to reproduce this visual structure; the result is that many of the fragments that we get are meaningless. We tried a very naive approach, filtering and dropping snippets based on a simple heuristics, but we want to test if this can be a important element to improve results.

Information is replicated across the web, and so we have fragments of different pages that have the same information. However, as visual presentation is not always the same, the results of the conversion to plain text produces different strings. We used the Dice Coefficient as measure to compare snippets and discover duplicates and almost duplicates.

On the other hand, one may wonder whether the retrieval of fragments in different languages provide more relevant information, and to what extent. These fragments in other languages are derived from queries which include terms in those other languages. We made a run using queries in English only, which should allow us to compare results and assess whether the extent to which the use of other languages aid in retrieval.

Cross-Language Geographical Retrieval (GeoCLEF)

GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview

Thomas Mandl[1], Paula Carvalho[2], Fredric Gey[3], Ray Larson[3], Diana Santos[2] and Christa Womser-Hacker[1]
[1] Information Science, University of Hildesheim, GERMANY
mandl@uni-hildesheim.de, womser@uni-hildesheim.de
[2] Linguateca, SINTEF ICT, NORWAY
Diana.Santos@sintef.no
[3]University of California, Berkeley, CA, USA
gey@berkeley.edu , ray@sims.berkeley.edu
WITH
Giorgio Di Nunzio[4], Nicola Ferro[4]
[4]Department of Information Engineering, University of Padua, Italy
{dinunzio|ferro}@dei.unipd.it

GeoCLEF is an evaluation initiative for testing queries with a geographic specification in large set of text documents. GeoCLEF ran a regular track for the third time within the Cross Language Evaluation Forum (CLEF) 2008. The purpose of GeoCLEF is to test and evaluate cross-language geographic information retrieval (GIR). GeoCLEF 2008 consisted of two sub tasks. A search task ran for the third time and a Wikipedia pilot task (GiKiP) was organized for the first time. For the GeoCLEF 2008 search task, twenty-five search topics were defined by the organizing groups for searching English, German and Portuguese document collections. Topics were developed also for English, German and Portuguese. Many topics were geographically challenging. Eleven groups submitted 131 runs. The groups used a variety of approaches, including sample documents, named entity extraction and ontology based retrieval. The evaluation methodology and results are presented in the paper.

Cheshire at GeoCLEF 2008: Text and Fusion Approaches for GIR

Ray R. Larson
University of California, Berkeley, School of Information, USA
ray@sims.berkeley.edu

In this paper we will briefly describe the approaches taken by Berkeley for the main GeoCLEF 2008 tasks (Monolingual and Bilingual retrieval). The approach this year used probabilistic text retrieval based on logistic regression and incorporating blind relevance feedback for all of the runs and in addition we ran a number of tests combining this type of search with OKAPI BM25 searches using a fusion approach. All translation for bilingual tasks was performed using the LEC Power Translator PC-based MT system.

Our results were good overall with Cheshire systems runs appearing in the top 5 participants for each task (German, English and Portuguese both Monolingual and Bilingual) with the highest ranked runs for Monolingual Portuguese and for Bilingual German, English and Portuguese. All of these top-ranked runs used the fusion approach.

However, once again this year we did not attempt to do any specialized geographic processing, because it appears that purely textual approaches to GIR are more effective when only textual topics, lacking explicit geographic coordinate constraints, are used.

Multi-lingual Geographical Information Retrieval

Rocio Guillen
California State University San Marcos
rguillen@csusm.edu

This paper reports on the results of our experiments in the Monolingual English, German and Portuguese tasks and the Bilingual German topics on English collections, English topics on German collections and English topics on Portuguese collections tasks.

Seven runs were submitted as official runs, four for the monolingual task and three for the bilingual task.

We used the Terrier (TERabyte RetrIEveR) Information Retrieval Platform version 2.1 to index and query the collections. Experiments were performed for both tasks using the Inverse Document Frequency model with Laplace after-effect and normalization 2. Topics were processed automatically and the only fields considered were the title and the description. We included the title field only for an experiment with the Portuguese collection. The stopword list provided by Terrier was used to index all the collections. Results for both the monolingual and bilingual tasks were low in terms of precision and recall mainly due to the following reasons:

1) no manual processing was done;

2) no query expansion based on automated relevance feedback was added;

3) no experiments including the narrative field were run;

4) no terms were translated for the bilingual task;

5) no German and Portuguese stopword lists were used instead of the default stopword list; and

6) no pre-processing or removal of diacritic marks was performed.

We are running new experiments to address some of the issues aforementioned and determine the impact they have on retrieval performance.

MMIS at GeoCLEF 2008: Experiments in GIR

Simon Overell[1], Adam Rae[2] and Stefan Rüger[2,1]
[1] Multimedia & Information Systems
Department of Computing, Imperial College London, SW7 2AZ, UK
[2] Knowledge Media Institute
The Open University, Milton Keynes, MK7 6AA, UK
seo01@doc.ic.ac.uk and {a.rae, s.rueger}@open.ac.uk

In this paper, we present our Geographic Information Retrieval System, Forostar, and the results of three experiments. We compare two data fusion methods, and show that a simple geographic filter outperforms a penalty based system. We compare context-based disambiguation to a default gazetteer and show no significant difference. Finally, we compare a unique geographic index to an ambiguous geographic index. The ambiguous index outperformed all other methods and was statistically significantly better than the baseline.

University of Hagen at GeoCLEF 2008: Combining IR and QA 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

This paper describes the participation of GIRSA at GeoCLEF 2008, the geographic information retrieval task at CLEF. GIRSA is a modified and improved variant of the system which participated at GeoCLEF 2007. It combines results retrieved with methods from information retrieval (IR) on geographically annotated data and question answering (QA) employing query decomposition.

For the monolingual German experiments, several parameter settings were varied: using a single index or a separate index for content and geographic annotation, using complex term weighting, adding location names from the narrative part of the topics, and merging results from IR and QA. The best mean average precision (MAP) was obtained by combining IR and QA results (0.2608 MAP).

For bilingual (English-German and Portuguese-German) experiments, topics were translated via various machine translation web services: Applied Language Solutions, Google Translate, and Promt Online Translator. Performance for these experiments is generally lower than for monolingual experiments. For both source languages, Google Translate seems to return the best translations. For English topics, 60% (0.1571 MAP) of the maximum MAP for monolingual German experiments is achieved. For bilingual Portuguese-German experiments, 80% (0.2085 MAP) of the maximum MAP for monolingual German experiments is achieved.

University of Pittsburgh at GeoCLEF 2008: Towards Effective Geographic Information Retrieval

Qiang Pu[1], Daqing He[2] and Qi Li[2]
[1]School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu, CHINA
puqiang@uestc.edu.cn
[2]School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, USA
{dah44,qil14}@pitt.edu

This paper reports University of Pittsburgh's participation in GeoCLEF 2008. As the first time participants, we only worked on the monolingual GeoCLEF evaluation. We developed two different methods for extracting geographic location information for query expansion. The first one is Geographic Information Retrieval with Geographic Coordinates Extraction and Clustering (GCEC). Its basic idea is that those locations in the same cluster with the original geographic location should be treated as the geographic approximations of the location which can be used for geographic query expansion. The second method is Wikipedia-based Geographic Information Retrieval (WIKIGEO). Geographic location names were mined from Wikipedia - the online encyclopedia which provides abundant types of knowledge. We also assume that a query in our geographic information retrieval task can be segmented into a topic part, a geo part and the relation part that separate the topic part from the geo part.

The following resources were employed in our tasks: 1) Indri version 2.4[1] was used as our information retrieval system; 2) LingPipe[2] was used as the named entity identification tool to markup queries, related articles and other extracted Web or Wikipedia information; 3) an online Chinese-English dictionary[3] was utilized for extracting synonyms for the topic part of the queries; 4) Google search engine was used to return top 10 retrieval results for expanding the geo part of the queries; 5) Wikipedia[4] was used to mine geographic location names as query expansion and be of the entrance to obtain geographic coordinates of a location.

Our experiments results show that: 1) our online geographic coordinate extraction and clustering algorithm is useful for the type of locations that do not have clear corresponding coordinates; 2) the expansion based on the geo-locations generated by GCEC is effectiveness in improving Geographic retrievals. 3) Using Wikipedia we can find the coordinates for many geo-locations, but its usage for query expansion still need further studies. 4) query expansion based on title only obtained better results than using the combination of title and narrative parts, which are thought to contain more related geographic information. Further study is need for this part too.

[1] http://www.lemurproject.org/indri/, Indri is a new search engine from the Lemur project; a cooperative effort between the University of Massachusetts and Carnegie Mellon University to build language modelling information retrieval tools.

[2] http://alias-i.com/lingpipe/, Natural language processing software for text analytics, text data mining and search.

[3] http://dict.cn/, an online Chinese-English dictionary.

[4] http://www.wikipedia.org/, an open and online encyclopedia written collaboratively by volunteers around the world and organized the knowledge in encyclopedia style.

SINAI-GIR System. University of Jaén at GeoCLEF 2008

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega and L. Alfonso Ureña-López
SINAI Research Group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,magc,mgarcia,laurena}@ujaen.es

In the third participation of the SINAI research group in GeoCLEF track, we have tried to improve the system proposed last year in GeoCLEF 2007. The main developments are related to the use of query reformulation, keywords recognition, hyponyms extraction and query geo-expansion. On the other hand, new rules have been applied in the Validator subsystem in order to filter the documents recovered by the IR subsystem. We employ resources such as the Geonames gazetteer, the Lemur toolkit as index-search engine or Lingpipe as entities recognizer. We have run a total of 15 experiments, combining these developments in order to resolve the monolingual and bilingual tasks. The results obtained shown that filtering does not reach yet to improve the baseline case. However, the use of keywords and hyponyms in the re-ranking process seems to improve the filtering results. Instead, the use of hyponyms does not improve in any case the results. On the other hand, the use of query reformulation and geo-expansion does not improve the baseline case either. Surprisingly we have obtained best results using only the content of the title and description labels from the topics (TD), unlike what happened in the 2007 experiments, where we reached the best results using the content of all labels (TDN).

The UPV at GeoCLEF 2008: The GeoWorSE System

Davide Buscaldi and Paolo Rosso
Natural Language Engineering Lab,
Dpto. Sistemas Informáticos y Computación (DSIC),
Universidad Politécnica de Valencia
{dbuscaldi, prosso}@dsic.upv.es

We participated to the English monolingual task of GeoCLEF 2008 with a new version of our 2007 system. This year it has been complemented with a map-based filter. During the indexing phase, all places are disambiguated and assigned their coordinates on the map. These coordinates are stored in a separate index. The search process is carried out in two phases: in the first one, we search the collection with the same method applied in 2007, which exploits the expansion of index terms by means of WordNet synonyms and holonyms. The next phase consists in a re-ranking of the results of the previous phase depending on the distance of document toponyms from the toponyms in the query, or depending on the fact that the document contains toponyms that are included in an area defined by the query. The area is calculated from the toponyms in the query and their meronyms. In this approach we use for the first time the GeoWordNet resource that allows to assign geographical coordinates to the places listed in WordNet. The obtained results show that the map-based filtering allows to improve the results over the base system, which uses only the textual information. The best result (25.4% in Mean Average Precision) was obtained with the filtering method and title and description only fields.

TextMESS at GeoCLEF 2008: Result Merging with Fuzzy Borda Ranking

José M. Perea-Ortega[1], L. Alfonso Ureña-López[1], Davide Buscaldi[2] and Paolo Rosso[2]
[1]Departamento de Informática, Universidad de Jaén
[2]Natural Language Engineering Lab
Dpto. Sistemas Informáticos y Computación (DSIC)
Universidad Politécnica de Valencia
{jmperea, laurena}@ujaen.es

This paper describe the joint participation by the Universidad Politécnica de Valencia and the Universidad of Jaén to the GeoCLEF English monolingual task. This activity has been carried out within the framework of the Spanish TextMESS project (Intelligent, Interactive and Multilingual Text Mining based on Human Language Technologies). The method employed for the participation is a result merging algorithm based on the fuzzy Borda voting scheme. This method takes as input the two document lists returned by the two systems developed by the participating groups and creates a document list where the documents are ranked according to the fuzzy Borda voting scheme. The results obtained are better than the individual systems, and also ones of the best ones of the task (28.4% in Mean Average Precision). However, the best result was obtained with a run which combined the baseline systems. The analysis of the results showed that the best runs were those in which only title and description were used, and unfortunately we chose to submit only a run of this type, with the base systems. The results confirm the effectiveness of the fuzzy Borda scheme for the combination of different systems: we obtained always an improvement over the performance of the two integrated systems in all the nine runs we submitted, even in cases where a system was sensibly worse than the other one.

INAOE at GeoCLEF 2008: A Ranking Approach Based on Sample Documents

Esaú Villatoro-Tello, Manuel Montes-y-Gómez and Luis Villaseor-Pineda
Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica Óptica y Electrónica, (INAOE), México
{villatoroe, mmontesg, villasen}@ccc.inaoep.mx

This paper describes the system developed by the Language Technologies Laboratory of INAOE for the Geographical Information Retrieval task of CLEF 2008. The presented system focuses on the problem of ranking documents in accordance to their geographical relevance. It is mainly based on the following hypotheses:

i) current IR machines are able to retrieve relevant documents for geographic queries, but they cannot generate a pertinent ranking; and

ii) complete documents provide more and better elements for the ranking process than isolated query terms.

Based on these hypotheses, our participation at GeoCLEF 2008 aimed to demonstrate that using some query-related sample texts it is possible to improve the final ranking of the retrieved documents. Experimental results indicated that our approach could improve the MAP (up to 0.318) of some sets of retrieved documents using only an average of two sample texts. These results also showed that the proposed approach is very sensitive to the presence of irrelevant sample texts as well as to the ambiguity of geographical terms.

Ontology-based Query Construction for GeoCLEF

Rui Wang[1] and Günter Neumann[2]
[1] Saarland University
66123 Saarbrücken, Germany
rwang@coli.uni-sb.de
[2] LT-Lab, DFKI
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
neumann@dfki.de

This paper describes our participation in GeoCLEF. Being different from the traditional information retrieval, we focus more on the query expansion instead of document ranking. We parse each topic into the event part and the geographic part and use different ontologies to expand both parts respectively. In all, we submit 5 runs and achieve 33.38% (best R-Prec) and 30.37% (best MAP) for the manual submissions and 33.19% and 29.24% for the automatic submissions, which show great advantages of our strategy for this task.

The University of Lisbon at GeoCLEF 2008

Nuno Cardoso, Patrícia Sousa and Mário J. Silva
University of Lisbon, Faculty of Sciences, LaSIGE
{ncardoso,csousa}@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt

We participated in GeoCLEF with the purpose of maturing the ideas first coined in last year's participation, given the encouraging results despite some observed limitations that we now want to address. The improvements have been made at three levels:

Query Processing: Our GIR methodology was initially cast on the assumption that the thematic and geographic facets of documents and queries were complementary and non-redundant. As the results failed to support this assumption, we questioned whether this segregational approach for GIR is indeed a good practice. This year, we experimented a new query processing approach, where geographic and non-geographic terms are not split into two independent classes, but used both as geographic feature selection criteria and plain query terms.

Text mining: Our shallow text mining approach often failed to capture essential geographic information to geo-reference documents. We developed a new named entity recognition module, REMBRANDT, and used it as a text annotation tool to recognise all kinds of named entities in the CLEF collection. We now generate more comprehensive geographic document signatures, which include two kinds of geographic features: i) explicit geographic evidence, consisting of grounded placenames that designate geographic locations, such as countries, divisions or territories, and ii) implicit geographic evidence, consisting of other grounded entities that do not explicitly designate geographic locations but are strongly related to a geographic location, such as monuments, buildings, company headquarters or summits.

Document Processing: We needed a simple ranking model that elegantly combined the text and geographic subspace models, eliminating the need for merging text and geographic ranking scores. We extended MG4J to suit this need by implementing an optimised BM25 weighting scheme over three index fields: i) text, for standard term indexes, ii) explicit local, for geographic terms considered as explicit geographic evidence, and iii) implicit local, for geographic terms associated to the implicit geographic evidence.

Our experiments aimed at: i) evaluating if the retrieval performance of the GIR system with local index fields outperforms standard text retrieval; ii) measuring the effect of considering the extracted implicit geographic evidence on retrieval results.

The obtained results, which are close to the best observed MAP values, show that our improved GIR system is consistently more effective when using the geographic signatures of the documents, outperforming BM25 retrieval in all GeoCLEF evaluation results since 2006.

GikiP

Getting Geographical Answers from Wikipedia: the GikiP Pilot at CLEF

Diana Santos[1], Nuno Cardoso[2], Paula Carvalho[2], Iustin Dornescu[3], Sven Hartrumpf[4], Johannes Leveling[4] and Yvonne Skalban[3]

[1]Linguateca, Oslo node, SINTEF ICT (Norway)

Diana.Santos@sintef.no

[2]University of Lisbon, DI, LasiGE, XLDB, Linguateca (Portugal)

ncardoso@xldb.di.fc.ul.pt, pqfcarvalho@gmail.com

[3]Research Group in Computational Linguistics (CLG) at the University of Wolverhampton (UK)

i.dornescu2@wlv.ac.uk, yvonne.skalban@wlv.ac.uk

[4]Intelligent Information and Communication Systems (IICS), University of Hagen (FernUniversität in Hagen) (Germany)

Sven.Hartrumpf@FernUni-Hagen.de, Johannes.Leveling@FernUni-Hagen.de

This paper reports on the GikiP pilot that took place in 2008 in GeoCLEF. The task, of retrieving the answer to open list questions with geographic information in Wikipedia in English, German and Portuguese, was conceived as a merge of QA and GIR in a realistic cross-lingual setup. For 15 topics, chosen so as to provide a balance between the three languages, participating systems had to send a list of answers in them, and were strongly encouraged to provide answers in more than one language.

Results were encouraging, both by showing that the task was interesting for humans and by demonstrating that there were already systems which could solve it in a satisfying way.

Three systems participated, two fully automatic (GIRSA-WP and WikipediaListQA@wlv) and one interactive (RENOIR) -- in which the user chose the set of smaller automatic procedures to be used. All systems tried to make use of relevant properties of Wikipedia, such as categories and cross-lingual alignment. Interestingly, the systems outperformed human performance, for two topics at least, which shows that the automation of the task is required.

Out of 662 answers by the systems, 179 were considered correct by the assessors, who found intriguing issues to be dealt with in future editions: the need for a better presentation format; the need to decide when information in different languages differ (or is even contradictory); and the need to further interact with the user to make the questions more precise.

The paper describes in detail the three participating systems. Also, it presents the topics and the criteria for their topic, the translation and assessment difficulties involved, and offer some ideas for improvements for future editions of GikiP or similar evaluation contests.

Cross-Language Video Retrieval (VideoCLEF)

Overview of VideoCLEF 2008: Automatic Generation of Topic-based Feeds for Dual Language Audio-Visual Content

Martha Larson[1], Eamonn Newman[2] and Gareth Jones[2]
[1]ISLA, University of Amsterdam
[2]CDVP School of Computing, Dublin City University
m.a.larson@uva.nl, {eamonn.newman | gareth.jones}@computing.dcu.ie

The VideoCLEF track, introduced in 2008, aims to develop and evaluate tasks related to analysis of and access to multilingual multimedia content. In its first year, VideoCLEF piloted the Vid2RSS task, which involved the classification of Dutch-language documentaries having embedded English content arising from interviews and discussions with non-Dutch speakers. Task participants were supplied with Dutch archival metadata, Dutch speech transcripts, English speech transcripts and 10 thematic category labels, which they were required to assign to the test set videos. Participants collected their own training data. Results were delivered in the form of a series of RSS-feeds, one for each category. Feed generation, intended to promote visualization, involved simple concatenation of existing feed items (title, description, keyframe). In addition to the main classification task, which was mandatory, VideoCLEF offered two discretionary tasks. The first was a translation task, requiring translation of the topic-based feeds from Dutch into a target language. The second was a keyframe extraction task, requiring selection of a semantically appropriate keyframe to represent the video from among a set of keyframes (one per shot) supplied with the test data. Five groups participated in the 2008 VideoCLEF track. The best runs produced f-scores higher than 0.50, although no group broke 0.60. A favorite strategy was to approach the task as a classification problem, collecting data from Wikipedia or using a general search engine to train classifiers (SVM, Naive Bayes and k-NN were used). A competitive approach was to treat the problem as an information retrieval task, with the class label as the query and the test set as the corpus. Both the Dutch speech transcripts and the archival metadata performed well as sources of indexing features, but no group succeeded in exploiting combinations of feature sources to significantly enhance performance. The translation task had one participant only, who translated the feeds to English. A small scale fluency/adequacy evaluation revealed the translation to be of sufficient quality to make it valuable to a non-Dutch speaking English speaker. The keyframe extraction test was performed also by only one participant, who deployed the strategy of selecting the keyframe from the shot with the most representative speech transcript content. The automatically selected shots were shown with a small user study to be competitive with manually selected shots. Future years of VideoCLEF will aim to expand the corpus and the class label list as well as to extend the track to additional tasks.

MIRACLE at VideoCLEF 2008: Classification of Multilingual Speech Transcripts

Julio Villena-Román[1,3] and Sara Lana-Serrano[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

This paper describes the participation of MIRACLE research consortium at the VideoCLEF (Vid2RSS) task at CLEF 2008. We have participated in the main mandatory Classification task that consists in classifying videos of television episodes using speech transcripts and metadata, and the Keyframe Extraction task, whose objective is to select keyframes that represent individual episodes from a set of supplied keyframes (one from each shot of the video source).

For the classification task, our system is composed of two main blocks. The first block is in charge of building a corpus that can be used as the core system knowledge base. The knowledge base for training the classifier was generated from Wikipedia articles. The second block contains the set of operational elements that are needed to classify the speech transcripts of the topic episodes and generate the output in RSS format. Those operational elements include an information retrieval system and a classifier, as well as modules for text extraction, filtering and RSS generation.

Our approach to keyframe extraction is based on the assumption that, in the context of a vector space model representation, the most representative fragment of each episode (represented by a vector) is the one whose distance to the whole episode (also a vector) is the lowest. After extracting the text from the episode transcription, the contents of both each shot and the whole episode are used to build a set of weighted vectors encoding the term frequency of the main most significant terms in the given episode. The keyframe extraction module selects the keyframe that belongs to the most representative shot in the episode, which is the shot whose vector has the lowest distance from (i.e., is the nearest to) the vector of the whole episode. The cosine distance is used for distance calculation.

We submitted different runs for each proposed subtask: three for the classification task and one for the keyframe extraction. For the classification task, the best micro-average precision (0.43) was achieved by the run in which only the Dutch transcription is used. When the knowledge base and the transcription in English are involved, results are significantly worse. This could be directly motivated by the fact that the dominant language of the episodes is Dutch. However, other possible explanations could be because the training set (the knowledge base) for English is much smaller than the one available for Dutch, or even because the voice recognition system for English is not as good as for Dutch. Anyway, comparing to other groups, we successfully ranked 3rd (out of 6 participants) in terms of precision and 2nd in terms of recall.

Regarding the keyframe extraction task, MIRACLE was the only participant who submitted results. Thus, the evaluation has been manually made comparing the automatic keyframe provided by our system against a manually selected one. On average, the subjects chose the automatic over the manually selected keyframe in 15.2 cases (41.08%) and the manually over the automatic in 21.8 cases (58.92%). Despite the subjectivity of this task and lack of any reference experiment to which compare our own system, these promising figures indicate that the automatically extracted keyframes may be strong competitors with the manual ones in the short- or middle-term future.

DCU at VideoCLEF 2008

Eamonn Newman and Gareth Jones
School of Computing, Dublin City University, Ireland
{eamonn.newman|gareth.jones}@computing.dcu.ie

We describe a baseline system for the VideoCLEF Vid2RSS task. For this task, systems were required to assign category labels to dual-language videos based on the ASR transcripts provided in the data set.

Our system uses an unaltered off-the-shelf implementation of Lucene as its base technology. Lucene is a software library which allows for straightforward construction of Information Retrieval systems for indexing and searching text.

The ASR content provided in the data set was indexed using Lucene's default stopword removal and tokenisation methods for both the English and Dutch content.

The subject categories were populated by using the category label as a query on the collection, and assigning the retrieved items to that category. These were then transformed to be publishable as RSS-feeds, which allowed for a simple visualisation of the results.

As this was a pilot task, no definitive conclusions could be reached from analysis of the results, but some areas for improving the system are identified.

SINAI at VideoCLEF 2008

José M. Perea-Ortega, Arturo Montejo-Ráez, M. Teresa Martín-Valdivia, Manuel C. Díaz-Galiano
and L. Alfonso Ureña-López
SINAI Research Group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,amontejo,maite,mcdiaz,laurena}@ujaen.es

This paper describes the first participation of the SINAI research group in the VideoCLEF 2008 track. We have only submitted runs for the classification task on Dutch and English languages. Our approach has consisted in the use of a particular Information Retrieval system as classification architecture, using the speech transcriptions as textual queries and generating textual corpus for each topic class. In order to generate this textual corpus we have used the Google search engine. We have employed Lemur as IR system, and the data has been preprocessed using the Dutch stemmer from Snowball for Dutch language and Porter stemmer for English. The experiments show that an IR system can perform well as classifier of multilingual videos, using their speech transcriptions and obtaining good results. Our results show that, despite the simplicity of our system, transcriptions are a good source of information for video classification. Anyhow, some enhancements on the system can be performed, by selecting additional sources of learning data such as Wikipedia.

VideoCLEF 2008: ASR Classification Based on Wikipedia Categories

Jens Kürsten, Daniel Richter and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media
09107 Chemnitz, Germany
[jens.kuersten j daniel.richter j maximilian.eibl] at cs.tu-chemnitz.de

This article describes our participation at the VideoCLEF track of the CLEF campaign 2008. We designed and implemented a prototype for the classification of the Video ASR data. Our approach was to regard the task as text classification problem. We used terms from Wikipedia categories as training data for our text classifiers. For the text classification the Naive-Bayes and kNN classifier from the WEKA toolkit were used. We submitted experiments for classification task 1 and 2. For the translation of the feeds to English (translation task) Google's AJAX language API was used. The evaluation of the classification task showed bad results for all of our experiments with a precision between 10 and 15 percent. These values did not meet our expectations. Interestingly, we could not improve the quality of the classification by using the provided metadata. But at least the created translation of the RSS Feeds was well.

The University of Amsterdam at VideoCLEF 2008

Jiyin He, Xu Zhang, Wouter Weerkamp and Martha Larson
ISLA, University of Amsterdam
{j.he,x.zhang,w.weerkamp,m.larson}@uva.nl

The University of Amsterdam (UAms) co-organized and participated in VideoCLEF track of CLEF 2008 in order to promote research in the area of analysis of multilingual audio and video and to further develop its own techniques for the classification and retrieval of conversational broadcast content.

The UAms team carried out the classification task, the primary sub-task of the Vid2RSS 2008 VideoCLEF task. This task involves the assignment of thematic category labels to dual language (Dutch/English) television episode videos. Videos receiving the same class labels are then depicted together in the form of a topic-based RSS-feed. UAms chose to focus on exploiting archival metadata and making use of speech transcripts generated by both the Dutch and English speech recognizers. Exploratory experimentation completed prior to the start of the task on external data motivated choosing a Support Vector Machine (SVM) with a linear kernel as the classifier. As a SVM toolbox to carry out the experiments, the Least Square-SVM (LS-SVM) toolbox was chosen. A significant challenge in the Vid2RSS 2008 task was that participants were required to collect their own training data. The UAms team chose to use data from Wikipedia because it contains material on the topics included in the list of Vid2RSS thematic categories and because it is multilingual.

The results of the task were less than satisfying and revealed that the exploitation of speech recognition transcripts for thematic classification of conversational broadcast content is far from a solved problem. The task results fail to demonstrate that satisfactory classification can be achieved using data collected from an independent source for training, but using speech recognition transcripts for classification. Moreover, no improvement was achieved by adding speech transcripts from the embedded language to the matrix language transcripts. The main positive result of the experiments was to demonstrate the potential of features derived from archival metadata in improving classification performance. In the case of a couple of the classes (music and history), the performance attained was nearly satisfactory. This result suggests that additional progress is achievable in the area of classifying video content in the face of a dearth of well-matched training data.

Further research is necessary, however, to gain an understanding of how to exploit speech transcripts, especially transcripts generated in the embedded language.

Multilingual Information Filtering (INFILE@CLEF)

Overview of CLEF 2008 INFILE Pilot Track

Romaric Besançon[1], Stéphane Chaudiron[2], Djamel Mostefa[3], Olivier Hamon[3,4], Ismaïl Timimi[2],
Khalid Choukri[3]
[1] CEA LIST
[2] Université de Lille 3 - GERiico
[3] ELDA
[4] LIPN (UMR 7030) – Université Paris 13 & CNRS

The INFILE campaign has been run for the first time as a pilot track in CLEF 2008. Its purpose is the evaluation of cross-language adaptive filtering systems. It uses a corpus of 300,000 newswires from Agence France Presse (AFP) in three languages: Arabic, English and French, and a set of 50 topics in general and specific domain (scientific and technological information).

The main features of the INFILE evaluation campaign are summarized here:

- Crosslingual: English, French and Arabic are concerned by the process but participants may be evaluated on mono or bilingual runs. - A newswire corpus provided by the Agence France Presse (AFP) and covering recent years (2004-2006).

- The topic set is composed of two different kinds of profiles, one concerning general news and events, and a second one on scientific and technological subjects.

- The evaluation is performed using an automatic interactive process for the participating systems to get documents and filter them, with a simulated user feedback.

- Systems are allowed to use the feedback at any time to increase performance.

- Systems provide a boolean decision for each document according to each profile.

- Relevance judgments are performed by human assessors.

The INFILE corpus is provided by the Agence France Presse (AFP) for research purpose. We selected 3 languages (Arabic, English and French) and a 3 years period (2004-2006) which represents a collection of about one and half millions newswires for around 10 GB, from which 100,000 documents of each language have been selected to be used for the filtering test.

A set of 50 profiles has been prepared covering two different categories. The first group (30 topics) deals with general news and events concerning national and international affairs, sports, politics, etc. The second one (20 topics) deals with scientific and technological subjects.

The results returned by the participants are binary decisions on the association of a document with a profile. Various metrics such as Precision, Recall, F-measure, linear utility, detection costs are used.

Only one participant actually submitted runs, the IMAG team, which submitted 3 runs, in monolingual English filtering.

IMAG obtained an F-measure of 0.36 (Precision 0.30 and recall 0.32).

Working Notes for the InFile Campaign: Online Document Filtering Using 1 Nearest Neighbor

Eric Gaussier, Ali Mustafa Qamar and Vincent Bodinier
Laboratoire d'Informatique de Grenoble (LIG)
Bâtiment IMAG B - 385 avenue de la Bibliothèque
38400 Saint Martin d'Hères
{eric.gaussier,ali-mustafa.qamar,vincent.bodinier}@imag.fr

This paper has been written as a part of the InFile (INformation, FILtering, Evaluation) campaign. This project is a cross-language adaptive filtering evaluation campaign, sponsored by the French national research agency, and it is a pilot track of the CLEF (Cross Language Evaluation Forum) 2008 campaigns. We propose in this paper an online algorithm to learn category specific thresholds in a multiclass environment where a document can belong to more than one class. Our method uses 1 Nearest Neighbor (1NN) algorithm for classification. It uses simulated user feedback to fine tune the threshold and in turn the classification performance over time. The experiments were run on English language corpus containing 100,000 documents. The best results have a precision of 0.366 and the recall is 0.260.

Morpho Challenge at CLEF 2008

Unsupervised Morpheme Analysis Evaluation by a Comparison to a Linguistic Gold Standard -- Morpho Challenge 2008

Mikko Kurimo and Matti Varjokallio
Adaptive Informatics Research Centre, Helsinki University of Technology
Mikko.Kurimo@tkk.fi

The goal of Morpho Challenge 2008 was to find and evaluate unsupervised algorithms that provide morpheme analyses for words in different languages.

Especially in morphologically complex languages, such as Finnish, Turkish and Arabic, morpheme analysis is important for lexical modeling of words in speech recognition, information retrieval and machine translation. The evaluation in Morpho Challenge competitions consisted of both a linguistic and an application oriented performance analysis.

This paper describes an evaluation where the competition entries were compared to a linguistic morpheme analysis gold standard. Because the morpheme labels in an unsupervised analysis can be arbitrary, the evaluation is based on matching the morpheme-sharing words between the proposed and the gold standard analyses. In addition to Finnish, Turkish, German and English evaluations performed in Morpho Challenge 2007, the competition this year had an additional evaluation in Arabic.

The results in 2008 show that although the level of precision and recall varies substantially between the tasks in different languages, the best methods seem to manage all the tested languages quite well.

The Morpho Challenge was part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF.

Unsupervised Morpheme Analysis Evaluation by IR Experiments -- Morpho Challenge 2008

Mikko Kurimo and Ville Turune
Adaptive Informatics Research Centre, Helsinki University of Technology
Mikko.Kurimo@tkk.fi

This paper presents the evaluation and results of Competition 2 (information retrieval experiments) in the Morpho Challenge 2008.

Competition 1 (a comparison to linguistic gold standard) is described in a companion paper.

In Morpho Challenge 2008 the goal was to search and evaluate unsupervised machine learning algorithms that provide morpheme analysis for words in different languages. The morpheme analysis can be important in several applications, where a large vocabulary is needed.

Especially in morphologically complex languages, such as Finnish, Turkish and Arabic, the agglutination, inflection, and compounding easily produces millions of different word forms which is clearly too much for building an effective vocabulary and training probabilistic models for the relations between words. The benefits of successful morpheme analysis can be seen, for example, in speech recognition, information retrieval, and machine translation.

In Morpho Challenge 2008 the morpheme analysis submitted by the Challenge participants were evaluated by performing information retrieval experiments, where the words in the documents and queries were replaced by their proposed morpheme representations and the search was based on morphemes instead of words.

The results indicate that the morpheme analysis has a significant effect in IR performance in all tested languages (Finnish, English and German). The best unsupervised and language-independent morpheme analysis methods can also rival the best language-dependent word normalization methods.

The Morpho Challenge was part of the EU Network of Excellence PASCAL Challenge Program and organized in collaboration with CLEF.

Retrieval Experiments at Morpho Challenge 2008

Paul McNamee
Johns Hopkins University
Human Language Technology Center of Excellence
paul.mcnamee@jhuapl.edu

Morpho Challenge 2008 hosted an extrinsic evaluation of morphological analysis that explored whether unsupervised morphology induction could benefit information retrieval. This paper presents results in alternative methods for word normalization using test sets in 13 languages from the Cross-Language Evaluation Forum (CLEF) ad-hoc evaluations between 2002 and 2007. Preliminary results for the Morpho Challenge 2008 evaluation are available in just English, Finnish and German. These results appear to be consistent with the larger set of CLEF experiments we conducted. We found that: (1) rule-based stemming is effective in less morphologically complicated languages; (2) alternative methods for stemming such as unsupervised learning of morphemes and least common n-gram stemming are helpful; and, (3) full character n-gram indexing is the most effective form of tokenization in more morphologically complex languages.

We examined a variety of methods for lexical normalization, including no transfomation, a rule-based stemmer (Snowball), segments produced by the Morfessor algorithm, least common n-grams from input words (of lengths 4 and 5), and regular character n-grams (of lengths 4 and 5). The most effective technique was character n-gram indexing which achieved a relative gain of 18% in mean average precision over unlemmatized words. In Czech, Bulgarian, Finnish, and Hungarian gains of over 40% were observed. While rule-based stemming can be quite effective, such tools are not available in every language and even when present, require additional work to integrate with an IR system. When language-neutral methods are able to achieve the same, or better performance, their use should be seriously considered.

Morphological Induction Through Linguistic Productivity

Sarah A. Goodman
University of Maryland-College Park
sagoodm@umd.edu

The induction program we have crafted relies primarily on the linguistic notion of ‘productivity’ to find affixes in unmarked text and without the aid of prior grammatical knowledge. In doing so, the algorithm unfolds in two stages. It first finds seed affixes, to include infixes and circumfixes, by assaying the character of all possible internal partitions of all words in a small corpus no larger than 3,000 tokens. It then selects a small subset of these seed affixes by examining the distribution patterns of roots they fashion to, as demonstrated in a possibly larger second training file. Specifically, it hypothesizes that valid roots take a partially overlapping affix-set, and develops this conjecture into agendas for both feature-set generation and binary clustering. It collects feature sets for each candidate by what we term affix-chaining, delineating (and storing) a path of affixes joined, with thresholding caveats, via the roots they share. After clustering these resultant sets, the program yields two affix groups, an ostensibly valid collection and a putatively spurious one. It refines the membership of the former by again examining the quality of shared root distributions across affixes. This second half of the program, furthermore, is iterative. This fact is again based in productivity, as we ration that, should a root take one affix, it most likely takes more. The code therefore seeds a subsequent iteration of training with affixes that associate with roots learned during the current pass. If, for example, it recognizes *view* on the first pass, and *viewership* occurs in the second training file, the program will evaluate ‘-ership’, along with its mate ‘-er’, via clustering and root connectivity on the second pass. The results of this method are thus far mixed according to training file size. Time constraints imposed by shortcomings in the algorithm's code, have thus far prevented us from fully training on a large file. For Morpho Challenge 2008, not only did we only train on just 1-30% of the offered text, thereby saddling the stemmer with a number of Out Of Vocabulary items, but, we also divided that text into smaller parts, thereby, as the results show, omitting valuable information about the true range of affix distributions.

Allomorfessor: Towards Unsupervised Morpheme Analysis

Oskar Kohonen, Sami Virpioja and Mikaela Klami
Adaptive Informatics Research Centre, Helsinki University of Technology
{oskar.kohonen,sami.virpioja,mikaela.klami}@tkk.fi

Morphological analysis is crucial to many modern natural language processing applications, especially when dealing with morphologically rich languages. Consequently, there has been an increasing amount of research on the task of unsupervised segmentation of word forms into smaller useful units, i.e. morphs or morphemes. Ultimately, we would like to perform not morphological segmentation, but the more difficult task of morpheme analysis, where the aim is not only to segment the corpus word forms into subparts, but also to identify surface forms corresponding morphological labels. For this task, the phenomenon of allomorphy places limits on the quality of morpheme analysis achievable by segmentation alone.

Our unsupervised method, Allomorfessor, tries to discover common baseforms for allomorphs from an unannotated corpus. The method does not directly model the corpus, but the lexicon of word forms in the corpus. At its core, the model is a probabilistic context-free grammar. The terminal symbols of the grammar are units resembling linguistical morphemes, specifically root stems and affixes. We call the non-terminal symbols virtual morphs; they are units that have substructure. Compared to a successful segmentation method, Morfessor Baseline, we add the notion of mutation to model allomorphic variation. Each virtual morph splits into two parts, prefix morph and suffix morph, with a potential mutation which modifies the prefix morph, which is assumed to be the baseform of the virtual morph. The applied mutations can sequentially delete or substitute letters of the prefix morph, starting from its end.

We use Maximum a Posteriori estimation and a local, greedy search procedure to obtain the model parameters. The computationally most challenging task is to find a good set of candidate baseforms and the mutations that modify them to the analyzed surface morph. We restrict the baseforms to those that exist in the initial word list and test only the K nearest candidates.

We evaluated the method by participating in the Morpho Challenge 2008 competition 1, where automatic analyses of corpora in English, German, Turkish and Finnish are compared against a linguistic gold standard. Our method achieved high precision but low recall for all the four languages. In practice, low recall means that the method undersegments, i.e., the analyses is only partial and most of the linguistical morphemes are not found. Despite the current problems in the algorithm, we find the general approach to be promising and the problem worth further research.

ParaMor and Morpho Challenge 2008

Christian Monson, Jaime Carbonell, Alon Lavie and Lori Levin
Language Technologies Institute, Carnegie Mellon University
cmonson@cs.cmu.edu

ParaMor, our unsupervised morphology induction system performed well at Morpho Challenge 2008. When ParaMor's morphological analyses, which specialize at identifying inflectional morphology, are added to the analyses from the general purpose unsupervised morphology induction system, Morfessor, the combined system identifies the morphemes of all five Challenge languages at recall scores higher than those of any other system which competed in Morpho Challenge. In Turkish, for example, the recall of the ParaMor-Morfessor system, at 52.1%, is twice that of the next highest system that participated. These strong recall scores lead to F1 values for morpheme identification as high as or higher than those of any competing system for all the competition languages but English. Of the three language tracks of the task-based information retrieval (IR) evaluation of Morpho Challenge, the combined ParaMor-Morfessor system placed first at average precision in the English and German tracks. And in the German and Finnish tracks of the IR task, the ParaMor-Morfessor system outperformed the hand-built stemming package, Snowball.

Using Unsupervised Paradigm Acquisition for Prefixes

Daniel Zeman
Ústav formalin a aplikované lingvistiky
Iniverzita Karlova, Praha, Czechia
zeman@ufal.mff.cuni.cz

We describe a simple method of unsupervised morpheme segmentation of words in an unknown language. All what is needed is a raw text corpus (or a list of words) in the given language. The algorithm identifies word parts occurring in many words and interprets them as morpheme candidates (prefixes, stems and suffixes). There are two main phases: /morpheme learning/ and proper /morpheme segmentation./ In the first phase, we learn morpheme candidates and filter them until we get lists of known morphemes. In the second phase, we get back to the original words and use the morpheme lists for segmenting of the words into morphemes.

In Zeman (2007) we only were able to cut the word in two parts at most: the stem and the suffix. The main innovation over Zeman (2007) is the ability to learn prefixes. We propose two algorithms for prefixes. “Reversed word” method is just the stem-suffix algorithm applied to a reversed word. “Rule-based” method is a more conservative one: required properties are specified and all prefixes complying with the constraints are learned.

Two segmentation algorithms have been tested: a strict (precision-oriented) one, and one less strict. The paper reports on more experiments than have been included in the main Morpho Challenge competition. The combination of Zeman (2007) stem-suffix learning, the rule-based prefix learning and the less strict segmentation is currently the most successful one. Resulting F-score of morpheme labeling heavily depends on language, ranging from 0.23 (Arabic) to 0.50 (English).

The error analysis section shows how typos affect the results. The current algorithm cannot use word frequencies and has no means of identifying typos. Numerous examples from data are shown and other suggestions for future work are made.

References:

Daniel Zeman. 2007. Unsupervised Acquiring of Morphological Paradigms from Tokenized Text. In: Working Notes for the Cross Language Evaluation Forum (CLEF) 2007 Workshop, Budapest, Hungary. ISSN 1818-8044. Revised version to appear in C. Peters et al. (eds.): CLEF 2007, LNCS 5152, pp. 892-899, Springer-Verlag, Berlin / Heidelberg, Germany, 2008.

Participating Institutions

Participant	Country	AdHoc	Dom Spec	iCLEF	QA	Image CLEF	Web CLEF	Geo CLEF	Video CLEF	Morpho Chall.	INFILE
Bulgarian Acad. Sci.	Bulgaria				X						
Cal. State - San Marcos	USA							X			
Carnegie Mellon U.	USA									X	
CEA-LIST	France					X					
Charles University	Czech Rep.									X	
CWI	Netherlands					X					
DFKI-Artificial Intelligence	Germany				X			X			
Dublin City U.	Ireland				X				X		
Hungarian Acad. Sci.	Hungary					X					
IDIAP Research Inst-	Switzerland					X					
Imperial College	UK					X		X			
INAOE-Ist. Nac.Astro-fisica, Optica, Electro.	Mexico	X			X	X		X			
Indian Statistical Inst.	India	X									
INESC-ID – Spoken Languages Lab.	Portugal				X						
INESC-ID – Data Management & IR	Portugal	X									
Int. Inst. Information Tech. (IIIT-Hyderabad)	India			X							
IPAL/IMAG-CNRS (IR2)	France/Singapore	X				X					
IRIT/SIG Toulouse	France					X					
Johns Hopkins U.	USA	X								X	
Know-Center	Austria	X									
Lab. LIG	France					X					X
LIMSI-CNRS	France				X						
Linguateca-SINTEF	Norway				X						
LINA-Nantes	France				X						
LSIS-CNRS	France					X					
Macedonian & Slovenian U. Team	Macedonia/ Slovenia					X					
Manchester Metropolitan U.	UK			X							
Microsoft Asia	China					X					
MIRACLE – Daedalus & Madrid Univs.	Spain				X	X			X		
Nat. Inst.Informatics	Japan					X					
Nat. Inst. Health	USA					X					
Nat.Taiwan U.	Taiwan					X					
Open Text Corp	Canada	X									
Open University	UK					X					
Oregon Health & Sci. U.	USA					X					
Participant	Country	AdHoc	Dom Spec	iCLEF	QA	Image CLEF	Web CLEF	Geo CLEF	Video CLEF	Morpho Chall	INFILE

Participant	Country	AdHoc	Dom Spec	iCLEF	QA	Image CLEF	Web CLEF	Geo CLEF	Video CLEF	Morpho Chall	INFILE
Priberam Informatica	Portugal				X						
Research Inst. for AI	Romania				X						
RWTH Aachen-HLT.	Germany					X					
RWTH Aachen - Med.Inf.	Germany					X					
Swedish Institute for Computer Science	Sweden			X							
SYNAPSE	France				X						
Tech. U. Chemnitz	Germany	X	X		X	X			X		
Tech. U. Darmstadt	Germany		X								
Tech. U. Helsinki	Finland									X	
Tel Aviv U.	Israel					X					
Telecom, Paris Tech	France					X					
TextMess	Spain					X		X			
U. & U.Hospitals Geneva	Switzerland					X					
U. Aberta	Portugal				X						
U. Alicante - GPLSI	Spain				X						
U. Alicante – Software and Comp. Systems	Spain	X		X	X	X					
U. AI.I Cuza Iasi	Romania				X						
U. Amsterdam	Netherlands		X						X		
U. Banjaluka	Bosnia and Herzegovina					X					
U. Bari	Italy	X									
U. Basel	Switzerland					X					
U. Basque Country	Spain	X			X						
UC Berkeley	USA	X	X					X			
U. Complutense de Madrid	Spain	X									
U. Concordia –CLAC	Canada					X					
U. Cordoba	Argentina				X						
U. Evora	Portugal				X						
U. Federal do Rio Grande do Sul	Brasil	X									
U. Geneva	Switzerland	X	X			X					
U. Groningen	Netherlands				X						
U. Hagen	Germany				X			X			
U. Hildesheim	Germany										
U. Jaen	Spain	X			X	X		X	X		
U. Jean Monnet	France					X					
U. Karlsruhe	Germany	X									
U. Koblenz-Landau	Germany				X						
U. Lisbon	Portugal							X			
U. Makere	Uganda					X					
U. Maryland & US Gov.	USA									X	
U. Meiji	Japan					X					
Participant	Country	AdHoc	Dom Spec	iCLEF	QA	Image CLEF	Web CLEF	Geo CLEF	Video CLEF	Morpho Chall	INFILE

Participant	Country	AdHoc	Dom Spec	iCLEF	QA	Image CLEF	Web CLEF	Geo CLEF	Video CLEF	Morpho Chall	INFILE
U.Nacional Colombia	Colombia					X					
UNED-LSI	Spain					X	X				
U. Neuchatel	Switzerland	X	X								
U. Ottawa	Canada					X					
U. Padova	Italy			X							
U. Peking	China					X					
U. Pittsburg	USA							X			
UPMC-LIP6	France					X					
U. Politecnica Catalunya	Spain				X						
U. Politecnica Valencia	Spain				X			X			
U. Porto	Portugal				X						
U. Salamanca	Spain						X				
U. Sheffield	UK			X		X					
U. Tehran1, NLP	Iran	X									
U. Tehran2, IRDB	Iran	X									
U. Tehran3, Electrical Computing 1.	Iran	X									
U. Tehran4, NLPDB	Iran	X									
U.Tehran6-NLPDB2	Iran	X									
U. Twente	Netherlands	X					X				
U. Tilberg	Netherlands				X						
U. Waseda	Japan					X					
U.Wolverhampton	UK				X						
U. York	UK									X	
Xerox SAS (CACAO)	EU Project	X									
Xerox XRCE	France	X				X					
Participant	Country	AdHoc	Dom Spec	iCLEF	QA	Image CLEF	Web CLEF	Geo CLEF	Video CLEF	Morpho Chall	INFILE

Coordination

CLEF is coordinated by the Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa. The following institutions have contributed to the organisation of the different tracks of the CLEF 2008 campaign:

- Athena Research Center, Greece
- Business Information Systems, Univ. of Applied Sciences Western Switzerland, Sierre, Switzerland
- Centre for the Evaluation of Human Language and Multimodal Communication Technologies (CELCT), Trento, Italy
- Centruum vor Wiskunde en Informatica, Amsterdam, The Netherlands
- Computer Science Department, University of the Basque Country, Spain
- Computer Vision and Multimedia Lab, University of Geneva, Switzerland
- Database Research Group, University of Tehran, Iran
- Department of Computer Science, RWTH Aachen University, Germany
- Department of Computer Science and Information Systems, University of Limerick, Ireland
- 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 & Clinical Epidemiology, Oregon Health & Science Univ., USA
- Department of Medical Informatics, Aachen University of Technology (RWTH), Germany
- Department of Medical Informatics, University Hospitals and University of Geneva, Switzerland
- Evaluations and Language Resources Distribution Agency Sarl, Paris, France
- German Research Centre for Artificial Intelligence, DFKI, Saarbrücken,
- GESIS-IZ Social Science Information Centre, Bonn, Germany
- Information Science, University of Groningen, The Netherlands
- Institute of Computer Aided Automation, Vienna University of Technology, Austria
- Intelligent Systems Lab, University of Amsterdam, Netherlands
- Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Orsay, France
- Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, Madrid, Spain
- Linguateca, Sintef, Oslo, Norway
- Linguateca, CISUC, Department of Information Engineering, University of Coimbra, Portugal
- Linguateca, XLDB, LasiGE, Department of Information Engineering, University of Lisbon, Portugal
- Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
- Microsoft Research Asia
- National Institute of Standards and Technology, Gaithersburg MD, USA
- Research Computing Center of Moscow State University, Russia
- Romanian Institute for Computer Science, Romania
- School of Computer Science and Mathematics, Victoria University, Australia
- School of Computing, Dublin City University, Ireland
- TALP Research Center, Universitat Politécnica de Catalunya, Barcelona, Spain.
- UC Data Archive and School of Information Management and Systems, UC Berkeley, USA

CLEF Steering Committee

- Maristella Agosti, University of Padova, Italy
- Martin Braschler, Zurich University of Applied Sciences Winterhur, Switzerland
- Amedeo Cappelli, ISTI-CNR & CELCT, Italy
- Hsin-Hsi Chen, National Taiwan University, Taipei, Taiwan
- Khalid Choukri, Evaluations and Language resources Distribution Agency, Paris, France
- Paul Clough, University of Sheffield, UK
- Thomas Deselaers, RWTH Aachen University, Germany
- Giorgio Di Nunzio, University of Padova, Italy
- David A. Evans, Clairvoyance Corporation, USA
- Marcello Federico, ITC-irst, Trento, Italy
- Nicola Ferro, University of Padova, Italy
- Christian Fluhr, CEA-LIST, Fontenay-aux-Roses, France
- Norbert Fuhr, University of Duisburg, Germany
- Frederic C. Gey, U.C. Berkeley, USA
- Julio Gonzalo, LSI-UNED, Madrid, Spain
- Donna Harman, National Institute of Standards and Technology, USA
- Gareth Jones, Dublin City University, Ireland
- Franciska de Jong, University of Twente, The Netherlands
- Noriko Kando, National Institute of Informatics, Tokyo, Japan
- Jussi Karlgren, Swedish Institute of Computer Science, Sweden
- Michael Kluck, German Institute for International and Security Affairs, Berlin, Germany
- Natalia Loukachevitch, Moscow State University, Russia
- Bernardo Magnini, ITC-irst, Trento, Italy
- Paul McNamee, Johns Hopkins University, USA
- Henning Müller, University & University Hospitals of Geneva, Switzerland
- Douglas W. Oard, University of Maryland, USA
- Anselmo Peñas, LSI-UNED, Madrid, Spain
- Maarten de Rijke, University of Amsterdam, The Netherlands
- Diana Santos, Linguateca, 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

Acknowledgments

The support of all the data providers and copyright holders that have contributed to the creation of the CLEF test collections is gratefully acknowledged, and in particular:

- The Los Angeles Times, for the American-English newspaper collection.
- SMG Newspapers (The Herald) for the British-English newspaper collection.
- Le Monde S.A. and ELDA: Evaluations and Language resources Distribution Agency, for the French newspaper collection.
- Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections.
- Hypersystems Srl, Torino and La Stampa, for the Italian newspaper data.
- Agencia EFE S.A. for the Spanish news agency 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.
- Hamshahri newspaper and DBRG, Univ. Tehran, for the Persian newspaper data.
- Público, Portugal, and Linguateca for the Portuguese (PT) newspaper collection.
- Folha, Brazil, and Linguateca for the Portuguese (BR) newspaper collection.
- Tidningarnas Telegrambyrå (TT) SE-105 12 Stockholm, Sweden for the Swedish newspaper data.
- Schweizerische Depeschenagentur, Switzerland, for the French, German & Italian Swiss news agency data.
- Ringier Kiadoi Rt. (Ringier Publishing Inc.).and the Research Institute for Linguistics, Hungarian Acad. Sci. for the Hungarian newspaper documents.
- Sega AD, Sofia; Standart Nyuz AD, Sofia, 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
- AFP Agence France Presse for the English, French and Arabic newswire data used in the INFILE track
- The British Library, Bibliothèque Nationale de France and the Austrian National Library for the library catalog records forming part of The European Library (TEL)
- InformationsZentrum Sozialwissen-schaften, Bonn, for the GIRT social science database.
- SocioNet system for the Russian Social Science Corpora
- Institute of Scientific Information for Social Sciences of the Russian Academy of Science (ISISS RAS) for the ISISS database
- University and University Hospitals, Geneva, Switzerland and Oregon Health and Science University for the ImageCLEFmed Radiological Medical Database.
- Aachen University of Technology (RWTH), Germany, for the IRMA annotated medical images.
- 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.
- Mallinkrodt Institue of Radiology for permission to use their nuclear medicine teaching file.
- University of Basel's Pathopic 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.
- USC Shoah Foundation Institute, and IBM (English) and The Johns Hopkins University Center for Language and Speech Processing (Czech) for the speech transcriptions.
- The Institute of Sound and Vision, The Netherlands, for the English/Dutch videos, the University of Twente for the speech transcriptions, and Dublin City University for the shot segmentation.

Without their contribution, this evaluation activity would be impossible.