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

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

Department of Computer Science. University of Jaén, Jaén, Spain
fdofer,amontejo,magc,laurenag@ujaen.es

This year, we have participated on Ad-Hoc Robust Multilingual track with the aim to evaluate two issues of CLIR systems. Firstly, the paper describes the method followed for query expansion in a multilingual environment by using web search results provided by the Google engine in order to increment retrieval robustness. Unfortunately, the results obtained are disappointing. The second issue reported is relative to the robustness of several usual merging algorithms. We have found that 2-step RSV merging algorithms perform better than others algorithms when geometric precision is applied.

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

Thomas Mandl, René Hackl, Christa Womser-Hacker

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

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

Comparing the Robustness of Expansion Techniques and Retrieval Measures

Stephen Tomlinson

Hummingbird, Ottawa, Ontario, Canada
stephen.tomlinson@hummingbird.com
http://www.hummingbird.com/

Hummingbird participated in the monolingual (Bulgarian, French, Hungarian, Portuguese and English) and robust (Dutch, English, French, German, Italian and Spanish) information retrieval tasks of the Ad-Hoc Track of the Cross-Language Evaluation Forum (CLEF) 2006. In all 22 of our experiments with blind feedback (a technique known to impair robustness across topics), the mean scores of the Average Precision, Geometric MAP and Precision@10 measures increased (and most of these increases were statistically significant), implying that these measures are not suitable as robust retrieval measures. In contrast, we found that measures based on just the first relevant item, such as a Generalized Success@10 measure, successfully discerned some robustness gains, particularly the robustness advantage of expanding Title queries by using the Description field instead of blind feedback.

BRUJA System. The University of Jaén at the Spanish Task of CLEFQA 2006

Miguel A. García-Cumbreras, L.A. Ureña-López, Fernando Martínez Santiago, Jose M. Perea-Ortega
University of Jaén
fmagc,laurena,dofer,jmpereag@ujaen.es

The paper presents our first participation in the bilingual English-Spanish track at CLEF QA 2006. The Multilingual BRUJA system is presented, a Question Answering (QA) system that works with questions in several languages and also collections in several languages. The BRUJA system is currently in its first phase of develop, so we have only run one official experiment with questions into English and the collection into Spanish. The results obtained shown that the prototype and its answer extraction phase, have to be finished and improved. An overall accuracy of 20.53% in not a good result and the system is in progress.

Cross Lingual Question Answering using QRISTAL for CLEF 2006

Dominique Laurent, Patrick Séguéla, Sophie Nègre
Synapse Développement, Toulouse, France
{dlaurent, p.seguela, sophie.negre }@synapsefr.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 CLEF 2005 for monolingual task (FrenchFrench) and multilingual task (EnglishFrench and PortugueseFrench). This article describes the improvements of the system since last year. Then, it presents our benchmarked results for the CLEF 2006 campaign and a critical description of the system. Since Synapse Développement is participating to Quaero project, QRISTAL is most likely to be integrated in a mass market search engine in the forthcoming years.

DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track

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

The paper describes QUANTICO, a cross-language open domain question answering system for German and English. The main features of the system are: use of preemptive off-line document annotation with syntactic information like chunk structures, apposition constructions and abbreviation-extension pairs for the passage retrieval; use of online translation services, language models and alignment methods for the cross-language scenarios; use of redundancy as an indicator of good answer candidates; selection of the best answers based on distance metrics defined over graph representations. Based on the question type two different strategies of answer extraction are triggered: for factoid questions answers are extracted from best IR-matched passages and selected by their redundancy and distance to the question keywords; for definition questions answers are considered to be the most redundant normalized linguistic structures with explanatory role (i.e., appositions, abbreviation's extensions). The results of evaluating the system's performance by CLEF were as follows: for the best German-German run we achieved an overall accuracy (ACC) of 42.33% and a mean reciprocal rank (MRR) of 0.45; for the best English-German run 32.98% (ACC) and 0.35 (MRR); for the German-English run 17.89% (ACC) and 0.17 (MRR).

IPAL Knowledge-based Medical Image Retrieval in ImageCLEFmed 2006

Caroline Lacoste, Jean-Pierre Chevallet, Joo-Hwee Lim, Xiong Wei, Daniel Raccoceanu, Diem Le Thi Hoang, Roxana Teodorescu, Nicolas Vuillenemot

IPAL French-Singaporean Joint Lab (I2R, CNRS, NUS, UJF)
viscl, viscjp, joohwee, wxiong, visdaniel@i2r.a-star.edu.sg

The paper presents the contribution of IPAL group on the CLEF 2006 medical retrieval task (i.e. ImageCLEFmed). The main idea of our group is to incorporate medical knowledge in the retrieval system within a multimodal fusion framework. For text, this knowledge is in the Unified Medical Language System (UMLS) sources. For images, this knowledge is in semantic features that are learned from examples within structured learning framework. We propose to represent both image and text using UMLS concepts. The use of UMLS concepts allows the system to work at a higher semantic level and to standardize the semantic index of medical data, facilitating the communication between visual end textual indexing and retrieval. The results obtained with UMLS-based approaches show the potential of this conceptual indexing, especially when using a semantic dimension filtering, and the benefit of working within a fusion framework, leading to the best results of ImageCLEFmed 2006. We also test a visual retrieval system based on manual query design and visual task fusion. Even if it provides the best visual results, this purely visual retrieval provides poor results in comparison to the best textual approaches.

ImageCLEFphoto&med

SINAI at ImageCLEF 2006

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

University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información, Spain
fmcdiaz,magc,maite,amontejo,laurenag@ujaen.es

The paper describes SINAI team participation in the ImageCLEF campaign. The SINAI research group participated in both the ad hoc task and the medical task. The experiments accomplished in both tasks result from very different approaches. For the adhoc task the main IR system used is the same as that of the 2005 ImageCLEF adhoc task. The improvement of the adhoc system is a new Machine Translation system that works with several translators and implements several heuristics. We have participated in the English monolingual task and in six bilingual tasks for the languages: Dutch, French, German, Italian, Portuguese and Spanish. The results obtained shown that the English monolingual results are good (0,2234 is our best result) and there is a loss of precision with the bilingual runs and some languages like German or Spanish works better than others, because of the translations. For the medical task, this year we carried out new and very different experiments to imageCLEFmed2005 ones. First of all, we have processed the set of collections using Information Gain (IG) to determine which are the best tags that should be considered in the indexing process. These tags are those supposed to provide the most relevant and non-redundant information, and have been selected automatically according to our information-based strategy along with the data and relevance assessments from last year. This year, our goal was to analyze how tag selection may contribute to the quality of final results. In order to select reduced set of tags we have computed IG. 11 different collections were generated according to the percentage of tags with highest IG value. Finally, only results related to experiments with selections over the 20%, 30% and 40% of available tags were submitted, since they reported best performance on 2005 data. Experiments using only textual query and using textual mixing with visual query have been submitted. For visual query we have used the GIFT lists provide by the organization. Surprisingly, the system performs better on the text retrieval alone than mixed textual and visual retrieval. On the other hand, we try show that information filtering through tag selection using information gain improves retrieval results without the need of a manual selection, but the obtained results are no conclusive. Unfortunately, the results obtained are not as successful as desired. Due to a computing processing mistake all our mixed runs obtain the same results than the visual GIFT baseline (0.0467). At the moment of writing of the paper we are modifying our system in order to solve this problem.

Monolingual and Bilingual Experiments in GeoCLEF2006

Rocio Guillén
California State University San Marcos, USA
rguillen@csusm.edu

The paper presents the results of our initial experiments in the monolingual English, Spanish and Portuguese tasks and the Bilingual Spanish → English, Spanish → Portuguese, English → Spanish and Portuguese → Spanish tasks. Twenty runs were submitted as official runs, thirteen for the monolingual task and seven for the bilingual task. We used the Terrier Information Retrieval Platform to run experiments for both tasks using the Inverse Document Frequency model with Laplace after-effect and normalization experiments included topics processed automatically as well as topics processed manually. Manual processing of topics was carried out using gazetteers (Alexandria Digital Library, European Parliament and GEOnet Names Server), some of them containing translations in languages other than English, others containing the latitude, longitude and area which allow for semi-automated spatial analysis (proximity analysis). For the bilingual task we developed a component based on the transfer approach in machine translation. Topics were pre-processed automatically to eliminate stopwords. Then topics in the source language were translated to the target language. A major problem we detected after submitting our results was that we did not include the Spanish newspaper collection for the year 95 (EFE 95) for indexing and retrieval purposes. Therefore, the results of our experiments with Spanish for the monolingual and bilingual tasks were affected in terms of recall and precision. We are currently rerunning experiments with the full Spanish collection for the monolingual and bilingual task to obtain a more accurate evaluation of the retrieval performance.

UNSW at GeoCLEF 2006

You-Heng Hu, Linlin Ge
School of Surveying and Spatial Information Systems
University of New South Wales, Sydney, Australia
you-heng.hu@student.unsw.edu.au, l.ge@unsw.edu.au

The paper describes our participation in the GeoCLEF monolingual English task of the Cross-language Evaluation Forum 2006. Our retrieval system consists of four modules: the geographic knowledge base; the indexing module; the document retrieval module and the ranking module. The geographic knowledge base provides information about important geographic entities around the world and relationships among them. The indexing module creates and maintains textual and geographic indices for document collections separately. The Boolean model is used in the document retrieval module to retrieve documents that meet both textual and geographic criteria. The ranking module applied ranking functions that are learned using Genetic Programming to the retrieved results. Performance evaluation of the implementation of these system modules is the main objective of this study. The results of our experiments show that the geographic knowledge base, the indexing module and the retrieval module are useful for geographic information retrieval tasks, but the proposed ranking function learning method does not work well.

SINAI at GeoCLEF 2006: Expanding the Topics with Geographical Information and Thesaurus

Manuel García-Vega, Miguel A. García-Cumbreras, L. Alfonso Ureña-López, José M. Perea-Ortega
University of Jaén, Spain
{mgarcia,magc,laurena,jmperea}@ujaen.es

The paper describes the first participation of the SINAI (Intelligent Systems of Access Information) group of the University of Jaén in GeoCLEF 2006. We have developed a system made up of three main modules. The first one is the translation subsystem that works with queries into Spanish, Portuguese and Deutsche. The second one is the query expansion subsystem that integrates a Named Entity Recognizer, a Gazetteer, a Thesaurus expansion module and a Geographical information module. The last subsystem is the Information Retrieval module, that works with collections and queries into English, and returns the result file. We have made several runs, that combines these modules to resolve the monolingual and the bilingual tasks. The results obtained shown that the use of geographical and thesaurus information for query expansion does not improve the retrieval, but this is the first step to try to improve the system in the future.

University of Hagen at GeoCLEF 2006: Experiments with Metonymy Recognition in Documents

Johannes Leveling, Dirk Veiel

Intelligent Information and Communication Systems (IICS)
University of Hagen, Germany
{johannes.leveling, dirk.veiel}@fernuni-hagen.de

The paper describes the participation of the IICS group at the GeoCLEF task of the CLEF campaign 2006. We describe different retrieval experiments using a separate index for location names and identifying and indexing of metonymic location names differently. The setup of our GIR system is a modified variant of the system setup for GeoCLEF 2005. We apply a classifier for the identification of metonymic location names for preprocessing the documents. This classifier is based on shallow features only and was trained on manually annotated data from the German CoNLL-2003 Shared Task corpus for Language-Independent Named Entity Recognition and from a subset of the GeoCLEF newspaper corpus. After preprocessing, documents contain additional information for location names that are to be indexed separately, i.e. LOC (all location names identified from), LOCLIT (location names in their literal sense) and LOCMET (location names in their metonymic sense). To obtain an IR query from the topic title, description, and narrative, we employ two methods. In the first method, a semantic parser analyzes the query text and the resulting semantic net is transformed into database query. The second method uses a Boolean combination of a bag-of-words (consisting of topical search terms) with location names. The results of our experiments can be summarized as follows: In contrast to results of earlier experiments, excluding metonymic sense of location names does not significantly improve mean average precision (MAP). A more detailed analysis showed that for some topics, the precision increased. Our experiments show that the additional use of topic narratives decreases MAP. For almost all experiment with the topic narrative, lower values for MAP and for the number of relevant and retrieve documents were observed. However, query expansion and the use of separate indexes improves the performance of our GIR application.

R2D2 at GeoCLEF 2006: a Mixed Approach

[1]Manuel García-Vega, [1]Miguel A. García-Cumbreras, [1]L. Alfonso Ureña-López, [1]José M. Perea-Ortega, [1]F. Javier Ariza-López, [2]Oscar Ferrández, [2]Antonio Toral, [2]Zornitsa Kozareva, [2]Elisa Noguera, [2]Andrés Montoyo, [2]Rafael Muñoz, [3]Davide Buscaldi, [3]Paolo Rosso

[1]University of Jaén
{mgarcia,magc,laurena,jmperea,fjariza}@ujaen.es
[2]University of Alicante
{ofe,atoral,zkozareva,elisa,montoyo,rafael}@dlsi.ua.es
[3]Polytechnical University of Valencia
{dbuscaldi,prosso}@dsic.upv.es

The paper describes the participation of a mixed approach in GeoCLEF-2006. We have participated in Monolingual English Task and we present a joint work of three groups or teams belonging to project R2D2 1 with a new system, mixing the 3 individual systems of the teams.

University of Twente at GeoCLEF 2006: Geofiltered Document Retrieval

Claudia Hauff, Dolf Trieschnigg, Henning Rode

University of Twente
Enschede, The Netherlands
{hauffc,trieschn,rodeh}@ewi.utwente.nl

In this report we describe the approach of the University of Twente to the 2006 GeoCLEF task. It is based on retrieval by content and the subsequent filtering by geographical relevance utilizing a gazetteer. The results do not show an improvement in retrieval performance when taking geographical information into account.