Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources

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

IIT Kharagpur, India

{debasis.mandal,pratyushb}@gmail.com

{sandipan, mayank, sudeshna} @cse.iitkgp.ernet.in

This paper describes our experiment on two cross-lingual and one monolingual English text retrievals at CLEF1 in the ad-hoc track. The cross-language task includes the retrieval of English documents in response to queries in two most widely spoken Indian languages, Hindi and Bengali. For our experiment, we had access to a Hindi-English bilingual lexicon, 'Shabdanjali', consisting of approx. 26K Hindi words. But neither we had any effective Bengali-English bilingual lexicon nor any parallel corpora to build the statistical lexicon. Under this limited resources, we mostly depended on our phoneme-based transliterations to generate equivalent English query from Hindi and Bengali topics. We adopted Automatic Query Generation and Machine Translation approach for our experiment. Other language-specific resources included a Bengali morphological analyzer, a Hindi stemmer and a set of 200 Hindi and 273 Bengali stop-words. Lucene framework was used for stemming, indexing, retrieval and scoring of the corpus documents. The CLEF results suggested the need for a rich bilingual lexicon for CLIR involving Indian languages. The best MAP values for Bengali, Hindi and English queries for our experiment were 7.26, 4.77 and 36.49 respectively.

Robust and More

Report of MIRACLE team for the Ad-Hoc track in CLEF 2007

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

[1] Universidad Politécnica de Madrid

[2] Universidad Carlos III de Madrid

[3] DAEDALUS - Data, Decisions and Language, S.A.

{josemiguel.goni, josecarlos.gonzalez, sara.lana}@upm.es,

julio.villena@uc3m.es

This paper presents the 2007 MIRACLE's team approach to the AdHoc Information Retrieval track. The work carried out for this campaign has been reduced to monolingual experiments, in the standard and in the robust tracks. No new approaches have been attempted in this campaign, following the procedures established in our participation in previous campaigns. For this campaign, runs were submitted for the following languages and tracks:

- Monolingual: Bulgarian, Hungarian, and Czech.
- Robust monolingual: French, English and Portuguese.

There is still some room for improvement around multilingual named entities recognition.

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

Fernando Martínez-Santiago, Arturo Montejo-Ráez, Miguel A. García-Cumbreras

Department of Computer Science. University of Jaén, Jaén, Spain

{dofer,amontejo,magc}@ujaen.es

We have reported on our experimentation for the Ad-Hoc Robust track CLEF task concerning web-based query generation for English and French collections. We have continued the approach of the last year, although the model has been modified. Last year we used Google in order to expand the original query. This year we don't expand the query but we rather make a new query to be executed. Thus, we have to deal with two lists of relevant documents, one from each query. In order to integrate both lists of documents we have applied logistic regression merging solution. Obtained results are discouraging.

Answer Validation Exercise

UNED at Answer Validation Exercise 2007

Álvaro Rodrigo, Anselmo Peñas, Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos, UNED

{alvarory, anselmo, felisa}@lsi.uned.esAbstract text.

The objective of the Answer Validation Exercise (AVE) 2007 is to develop systems able to decide if the answer to a question is correct or not. Since it is expected that a high percent of the answers, questions and supporting snippets contain named entities, the paper presents a method for validating answers that uses only information about named entities. The promising results aim us to improve the system and use it as a component of other systems.

DFKI–LT at AVE 2007: Using Recognizing Textual Entailment for Answer Validation

Rui Wang, Günter Neumann

LT-lab, DFKI

Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

{wang.rui, neumann}@dfki.de

This report is about our participation in the *Answer Validation Exercise* (AVE) 2007. Our system utilizes a *Recognizing Textual Entailment* (RTE) system as a component to validate answers. We first change the question and the answer into *Hypothesis* (**H**) and view the document as *Text* (**T**), in order to cast the AVE task into a RTE problem. Then, we use our RTE system to tell us whether the entailment relation holds between the documents (i.e. **T**s) and question-answer pairs (i.e. **H**s). Finally, we adapt the results for the AVE task. In all, we have submitted two runs and achieved f-measures of 0.46 and 0.55 respectively, which both outperform last year's best result for English. After detailed error analysis, we have found that both the recall and the precision of our system could be improved in the future.

SINAI at QA@CLEF 2007. Answer Validation Exercise

M.A. García-Cumbreras, J. M. Perea-Ortega, F. Martínez Santiago, L.A. Ureña-López

University of Jaén. Computers Department

SINAI Group

Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain

{magc,jmperea,dofer,laurena}@ujaen.es

Information) group of the University of Jaén in the AVE task of QA@CLEF 2007. We have developed a system made up of training and classification processes, that uses machine learning methods (bbr, timbl). Based on lexical features it obtains good results, a 41% of QA accuracy.

ImageCLEFphoto&med

SINAI at ImageCLEF 2007

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

University of Jáen. Departamento de Informática

Grupo Sistemas Inteligentes de Acceso a la Información

Campus Las Lagunillas, Ed. A3, E-23071, Jáen, Spain

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

This paper describes the SINAI team participation in the ImageCLEF campaign. The SINAI research group has participated in both the ad hoc task and the medical task. The experiments accomplished in both cases result from very different approaches.

For the ad hoc task the main Information Retrieval (IR) system used combines the document lists retrieved by two IR systems, and uses online translators for the bilingual experiments. For the medical task, we have used the MeSH ontology to expand the queries. The expansion consists in searching terms of the query in the MeSH ontology in order to add similar terms. We have processed the set of collections using Information Gain (IG) in the same way as in ImageCLEFmed 2006.

University and Hospitals of Geneva at ImageCLEF 2007

Xin Zhou, Julien Gobeill, Patrick Ruch, Henning Müller

University and Hospitals of Geneva, Switzerland

xin.zhou@sim.hcuge.ch

This article describes the participation of the University and Hospitals of Geneva at three tasks of the 2007 ImageCLEF image retrieval benchmark. Two of these tasks were medical tasks and one was a photographic retrieval task. The visual retrieval techniques relied mainly on the GNU Image Finding Tool (GIFT) whereas multilingual text retrieval was performed by mapping the full text documents and the queries in a variety of languages onto MeSH (Medical Subject Headings) terms, using the EasyIR text retrieval engine for retrieval.

For the visual tasks it becomes clear that the baseline GIFT runs do not have the same performance as more sophisticated modern techniques do. GIFT can be seen as a baseline for the visual retrieval as it has been used for the past four years in ImageCLEF. Whereas in 2004 the performance of GIFT was among the best systems it now is towards the end of the spectrum, showing the clear improvement in retrieval quality of participants over the years. Due to time constraints no optimisations could be performed and no relevance feedback was used, usually one of the strong points of GIFT. The text retrieval runs have a fairly good performance showing the effectiveness of the approach to map terms onto an ontology. Mixed runs are in performance slightly lower than the best text results alone, meaning that more care needs to be taken in combining runs other than a simple linear combination. English is by far the language with the best results; even a mixed run of the three languages was lower in performance. This can partly be explained with the judges as they are all native English speakers. Thus, a bias towards relevance for English documents is unfortunately possible.

FIRE in ImageCLEF 2007

Thomas Deselaers, Tobias Gass, Tobias Weyand, Hermann Ney

Human Language Technology and Pattern Recognition Group

RWTH Aachen University, Aachen, Germany

deselaers@cs.rwth-aachen.de

We present the methods we applied in the four different tasks of the ImageCLEF 2007 content-based image retrieval evaluation. We participated in all four tasks using a variety of methods. Global and local image descriptors are applied using nearest neighbour search for the medical and photo retrieval tasks and discriminative models for the object retrieval and the medical automatic annotation task. For the photo and medical retrieval task, we apply a maximum entropy training method to learn an optimal feature weighting from the queries and qrels from last year. This method works particularly well if the queries are very similar as they were in the medical retrieval task.

SINAI at CL-SR task at CLEF 2007

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

University of Jaén. Departamento de Informática

Grupo Sistemas Inteligentes de Acceso a la Información

Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain

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

This paper describes the first participation of the SINAI team in the CLEF 2007 CLSR track. This year, we only want to establish a first contact with the task and the collections. Thus, we have pre-processed the collection using the Information Gain technique in order to filter the labels with most relevant information. We have used the LEMUR toolkit as the Information Retrieval system in our experiments.

Model Fusion Experiments for the Cross Language Speech Retrieval Task at CLEF 2007

Muath Alzghool, Diana Inkpen

School of Information Technology and Engineering

University of Ottawa

{alzghool, diana}@site.uottawa.ca

This paper presents the participation of the University of Ottawa group in the Cross-Language Speech Retrieval (CL-SR) task at CLEF 2007. We present the results of the submitted runs for the English collection. We have used two Information Retrieval systems in our experiments: SMART and Terrier, with two query expansion techniques: one based on a thesaurus and the second one based on blind relevant feedback. We proposed two novel data fusion methods for merging the results of several models (retrieval schemes available in SMART and Terrier). Our experiments showed that the combination of query expansion methods and data fusion methods helps to improve the retrieval performance. We also present cross-language experiments, where the queries are automatically translated by combining the results of several online machine translation tools. Experiments on indexing the manual summaries and keywords gave the best retrieval results.

Attempts to Search Czech Spontaneous Spoken Interviews - the University of West Bohemia at CLEF 2007 CL-SR track

Pavel Ircing, Luděk Müller

University of West Bohemia

{ircing, muller}@kky.zcu.cz

The paper presents an overview of the system build and experiments performed for the CLEF 2007 CL-SR track by the University of West Bohemia. We have concentrated on the monolingual experiments using the Czech collection only. The approach that was successfully employed by our team in the last year's campaign (simple tf.idf model with blind relevance feedback, accompanied with solid linguistic preprocessing) was used again but the set of performed experiments was broadened.

GEOUJA System. University of Jaén at GEOCLEF 2007

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, Arturo Montejo-Ráez

SINAI Group. Department of Computer Science. University of Jaén

Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain

{jmperea,magc,mgarcia,amontejo}@ujaen.es

This paper describes the second participation of the SINAI group of the University of Jaén in GeoCLEF 2007. We have developed a system different from the one presented in GeoCLEF 2006. Our architecture is made up of five main modules. The first one is the Information Retrieval Subsystem, that works with collections and queries in English and returns relevant documents for a query. The queries that are not in English are translated by the Translation Subsystem. All the queries are filtered by the Geo-Relation Finder Subsystem, that finds any spatial relation in the topic, and NER (Named Entities Recognition) Subsystem, that looks for any location in the topic. The most important module is the Geo-Relation Validator Subsystem, it applies some heuristics to filter documents recovered by the IR Subsystem. We have made several runs, combining these modules to resolve the monolingual and the bilingual tasks. The results obtained show that the heuristics applied are quite restrictive and therefore it must be generated new heuristics and to improve the definition of new rules to filter recovered documents.

Monolingual Retrieval Experiments with Spatial Restrictions at GeoCLEF 2007

Ralph Kölle, Ben Heuwing, Thomas Mandl, Christa Womser-Hacker

Information Science, University of Hildesheim,

Marienburger Platz 22

D-31141 Hildesheim, Germany

koelle@uni-hildesheim.de

The participation of the University of Hildesheim focused on the monolingual German and English tasks of GeoCLEF 2007. Based on the results of GeoCLEF 2005 and GeoCLEF 2006, the weighting and expansion of geographic named entities (NE) and Blind Relevance Feedback were combined. This year an improved model for German Named Entity Recognition was evaluated.

GeoCLEF2007 Experiments in Query Parsing and Cross-Language GIR

Rocio Guillén

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 Spanish ↦ English, Spanish → Portuguese tasks. We also present initial results on the recognition, extraction and categorization of web-based queries for the Query Parsing task. Twenty-three runs were submitted as official runs, 16 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 2 and the Ponte-Croft language model. Experiments included topics processed automatically as well as topics processed manually. Manual processing of topics was carried out for the bilingual task using the transfer approach in machine translation. Topics were pre-processed automatically to eliminate stopwords. Results show that automatic relevance feedback with 5 terms and 20 documents performs better, in general. The initial approach used in the Query Parsing task is a pattern-based approach. Due to the ungrammaticality, multilinguality and ambiguity of the language in the 800,000 web-based queries in the collection, we started by building a list of all the different words in the queries, similar to creating an index. Next, a lookup of the words was done in a list of countries to identify potential locations. Because many locations were missed, we further analyzed the queries looking for spatial prepositions and syntactic cues. Queries were processed by combining search in gazetteers with a set of patterns. Categorization was also based on patterns. Results were low in terms of recall and precision.