GEOUJA System. The first participation of the University of Jaén at GEOCLEF 2006

Manuel García-Vega[1], Miguel A. García-Cumbreras[1], L.A. Ureña-López[1] and José M. Perea-Ortega[1]

Dpto. Computer Science. University of Jaén.Spain
{mgarcia,magc,laurena,jmperea}@ujaen.es

Abstract. This paper describes the first participation of the SINAI group of the University of Jaén in GeoCLEF 2006. We have developed a system made up of three main modules: the Translation Subsystem, that works with queries into Spanish and German against English collection; the Query Expansion subsystem, that integrates a Named Entity Recognizer, a thesaurus expansion module and a geographical information-gazetteer module; and the Information Retrieval subsystem. We have participated in the monolingual and the bilingual tasks. The results obtained shown that the use of geographical and thesaurus information for query expansion does not improve the retrieval in our experiments.

1 Introduction

The objective of GeoCLEF is to evaluate Geographical Information Retrieval (GIR) systems in tasks that involves both spatial and multilingual aspects. Given a multilingual statement describing a spatial user need (topic), the challenge is to find relevant documents from target collections into English, using topics into English, Spanish, German or Portuguese [1, 2].

The main objective of our first participation in GeoCLEF have been the study of the problem of this task, and to develop a system to solve some of them. For this reason, our system consist of three subsystems: Translation, Query Expansion and Information Retrieval. The Query Expansion subsystem is formed as well by three modules: Named Entity Recognition, Geographical Information-Gazetteer and Thesaurus Expansion.

Next section describes the whole system and each module of the system. Then, in the section 3 experiments and results are described. Finally, the conclusions about our first participation in GeoClef 2006 are expounded.

2 System Description

We propose a Geographical Information Retrieval System that is made up of three subsystems (See Figure 1):

- **Translation Subsystem**: is the query translation module. This subsystem translates the queries to the other languages. For the translation an own module has been used, called SINTRAM (SINai TRAnslation Module)[1], that works with several online Machine Translators, and implements several heuristics. For these experiments we have used an heuristic that joins the translation of a default translator (the one that we indicate depends of the pair of languages), with the words that have another translation (using the other translators).
- **Query Expansion Subsystem**: the goal of this subsystem is to expand the queries with geographical data and words from the thesaurus (see below).It is made up of three modules: a Named Entity Recognition Module, that uses a Geographical Information-Gazetteer Module, and a Thesaurus Expansion Module. These modules are described in detail next:
  - **Named Entity Recognition (NER) Module**: the main goal of NER Subsystem is to detect and recognize the *location* entities in the queries, in order to expand the topics with geographical data (using the Geographical Information Module). We have used the NER module of GATE[2] and its own Gazetteer. The location terms includes everything that is town, city, capital, country and even continent. The NER Module generates some labelled topics adding the found locations.
  - **Geographical Information Module**: This module stores the geographical data. This information has been obtained from the Geonames database[3] . We have used it only for English because previously all the queries have been translated. The goal of this module is to expand the locations of the topics recognized by the NER module, using geographical information. The have made automatic query expansion[3]. When a location is recognized by the NER Module the system look for in the Geographical Information Module. In addition, it is necessary to consider the spatial relations found in the query ("near to", "within X miles of", "north of", "south of", etc.). Depending on the spatial relations, the search in the Geographical Information Module is more or less restrictive.
  - **Thesaurus Expansion Module**: This is the query expansion module, using an own thesaurus. A collection of thesauri was generated from the GeoCLEF English training corpus. This module was looking for words with a very high rate of document co-location, using the standard *TF.IDF* for words comparing test. These words were treated like synonyms and added to the topics. For that, we generated an inverse file with the GeoCLEF 2005 corpus. The file has a row for each different word of the corpus with the word frequencies for each corpus file. We found that a cosine similarity great than 0.9 between words was the rate

[1] http://sinai.ujaen.es
[2] http://gate.ac.uk/
[3] http://www.geonames.org/. Geonames geographical database contains over eight million geographical names and consists of 6.3 million unique features whereof 2.2 million populated places and 1.8 million alternate names

that obtain best precision/recall results (in average 2 words added). The same procedure was applied to the 2006 corpus.

– **Information Retrieval Subsystem**: We have used LEMUR IR system[4].

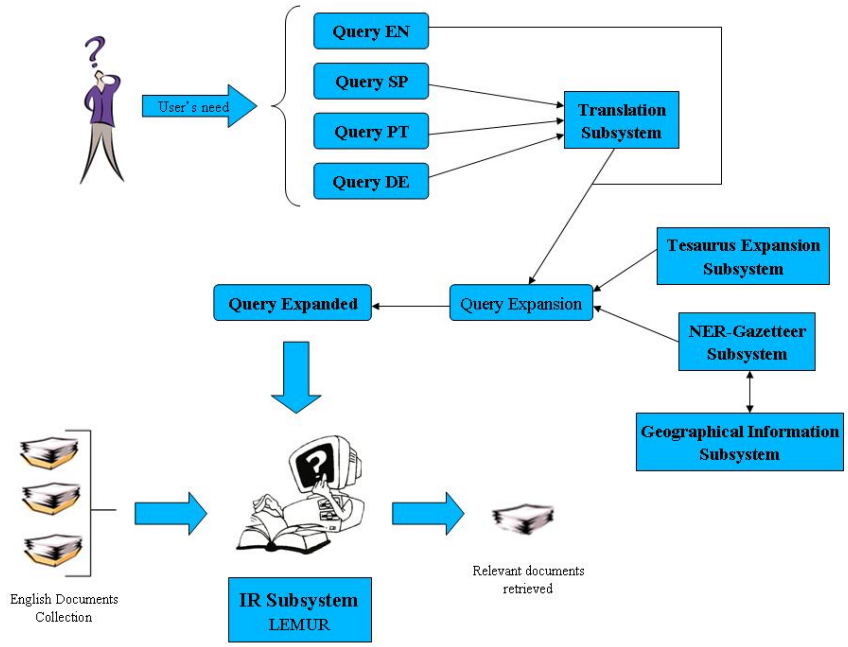

Fig. 1. GEOUJA System architecture

3 Results

In our **baseline** experiment we have used the original English topics set. All topics are preprocessed (stopper and stemmer) without query expansion (geographical or thesaurus). We have used Okapi as weighting function and pseudo-relevant feedback(PRF) in all experiments.

We have participated in monolingual and bilingual tasks. In monolingual task with five experiments (the official results are shown in Table 1):

– *sinaiEnEnExp1*. Baseline experiment, considering all tags (title, description and narrative). Without expansion topics.
– *sinaiEnEnExp2*. The same experiment as baseline but considering only title and description tags.

[4] http://www.lemurproject.org/

– *sinaiEnEnExp3*. We have considered only title and description tags. The query has been expanded using the title with geographical information.
– *sinaiEnEnExp4*. We have considered only title and description tags. The query has been expanded using the title and the description with thesaurus information.
– *sinaiEnEnExp5*. We have considered only title and description tags. The query has been expanded using the title and the description with geographical and thesaurus information.

For the bilingual task we have made five experiments: two experiments for German-English task and three experiments for Spanish-English task (the official results are shown in Table 2):

– *sinaiDeEnExp1*. All the query tags have been translated from German to English and preprocess (stopper and stemmer), and the query has not been expanded.
– *sinaiDeEnExp2*. The same experiment as previous but only considering the title and description tags.
– *sinaiEsEnExp1*. All the query tags have been translated from Spanish to English and preprocess (stopper and stemmer), and the query has not been expanded.
– *sinaiEsEnExp2*. The same experiment as previous but only considering the title and description tags.
– *sinaiEsEnExp3*. We have considered only title and description tags. The query has been expanded using the title with geographical information.

Experiment	Mean Average Precision	R-Precision
sinaiEnEnExp1	0.3223	0.2934
sinaiEnEnExp2	0.2504	0.2194
sinaiEnEnExp3	0.2295	0.2027
sinaiEnEnExp4	0.2610	0.2260
sinaiEnEnExp5	0.2407	0.2094

Table 1. Official results in monolingual task

Experiment	Mean Average Precision	R-Precision
sinaiDeEnExp1	0.1868	0.1649
sinaiDeEnExp2	0.2163	0.1955
sinaiEsEnExp1	0.2707	0.2427
sinaiEsEnExp2	0.2256	0.2063
sinaiEsEnExp3	0.2208	0.2041

Table 2. Official results in bilingual task

4 Conclusions

The obtained results shown that the way we have implemented query expansion (using geographical and thesaurus information) did not improve the retrieval. Several reasons exist to explain the worse results obtained with the expansion of topics, and these are our conclusions:

- The NER module used sometimes does not work well, because in some topics only a few entities are recognized and not all. For the future we will test other NERs.
- In the topics, sometimes appear compound locations like New England, Middle East, Eastern Bloc, etc., that not appear in the Geographical Information-Gazetteer Module.
- Depending on spatial relation in the topics, we could improve the expansion, testing in which cases the system works better with more locations or less. Therefore, we will try to improve the Geographical Information-Gazetteer Module and the Thesaurus Expansion Module to obtain better query expansions.

As future work we want to know why the expansion module did not work as well as we expected. It is known that sometimes the Geonames module introduce noise in the queries, but our thesauri should improve the baseline method. We also want to include in the system another module that expand the queries using Google.

5 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

References

1. Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, Henning Mller:Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks. In Working Notes for the CLEF 2006 Workshop, 2006.
2. Henning Mller, Thomas Deselaers, Thomas Lehmann, Paul Clough, Eugene Kim, William Hersh: Overview of the ImageCLEFmed 2006 Medical Retrieval and Annotation Tasks. In Working Notes for the CLEF 2006 Workshop, 2006.
3. D. Buscaldi and P. Rosso and E. Sanchis-Arnal: A WordNet-based Query Expansion method for Geographical Information Retrieval. In Working Notes for the CLEF 2005 Workshop, 2005.