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

Abstract

This 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.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Algorithms, Languages, Performance, Experimentation

Keywords

Information Retrieval, Geographic Information Retrieval, Named Entity Recognition, GeoCLEF

1 Introduction

This year is the fourth participation of the SINAI (Intelligent Systems of Access Information) group of the University of Jaén in CLEF, since 2001 [2, 5, 6, 3, 4]. The first years we participated in the Adhoc CLIR tasks, with good results, and each year we try to build other systems. In this campaign we participate in Cross-Language Information Retrieval, Question Answering, GeoCLEF and ImageCLEF.

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 in English, using topics in English, Spanish, German or Portuguese.

The main objective of our first participation in GeoCLEF have been the study of the problem of this task, and the development of a system that returns relevant documents.

The Cross-Language Geographic Information Retrieval (GIR) system developed at the University of Jaén has been designed to retrieve relevant documents that contain geographic tags.

For this reason, our system consist of several modules: Translation, Named Entity Recognition-Gazetteer, Geographical Information Subsystem and Thesaurus Expansion Subsystem.

This paper is organized as follows: section 2 describes the whole system and each module of the system in detail. Then, in the section 3 experiments and results are described.

Finally, the conclusions about our participation in GeoClef 2006 are expounded, and some future work.

2 System Description

We propose a Geographical Information Retrieval System that is made up of five subsystems:

- **Translation Subsystem**: is the query translation module. This subsystem translates the queries into the other languages.
- **Named Entity Recognition-Gazetteer Subsystem**: is the query geo-expansion module. This subsystem uses the geographical information from Geographical Information module.
- **Geographical Information Subsystem**: is the module that stores the geographical data. This information has been obtained from Geonames[1] gazetteer.
- **Thesaurus Expansion Subsystem**: is the query expansion module using an own Thesaurus.
- **IR Subsystem**: is the Information Retrieval module. We have used the LEMUR IR system[2].

The combination of these subsystems gives the results of the different runs, and will be shown in section 3.

2.1 Translation Subsystem

The purpose of the Translation Subsystem is to translate the queries or topics that are not in English.

This module is used for the following bilingual tasks: Spanish-English, Portuguese-English and German-English.

For the translation we have used an own module, called SINTRAM (SINai TRAnslation Module), that works with several online Machine Translators, and implements several heuristics. In this case we have used an heuristic that joins the translation of a default translator (the one that we indicate, depending of the pair of languages) with these words that have another translation (using another translator).

SINTRAM works with the following online translators:

- Epals: available at *http://www.epals.com*
- Prompt: available at *http://translation2.paralink.com*
- Reverso: available at *http://www.reverso.net*
- Systran: available at *http://www.systransoft.com*

[1] http://www.geonames.org
[2] http://www.lemurproject.org

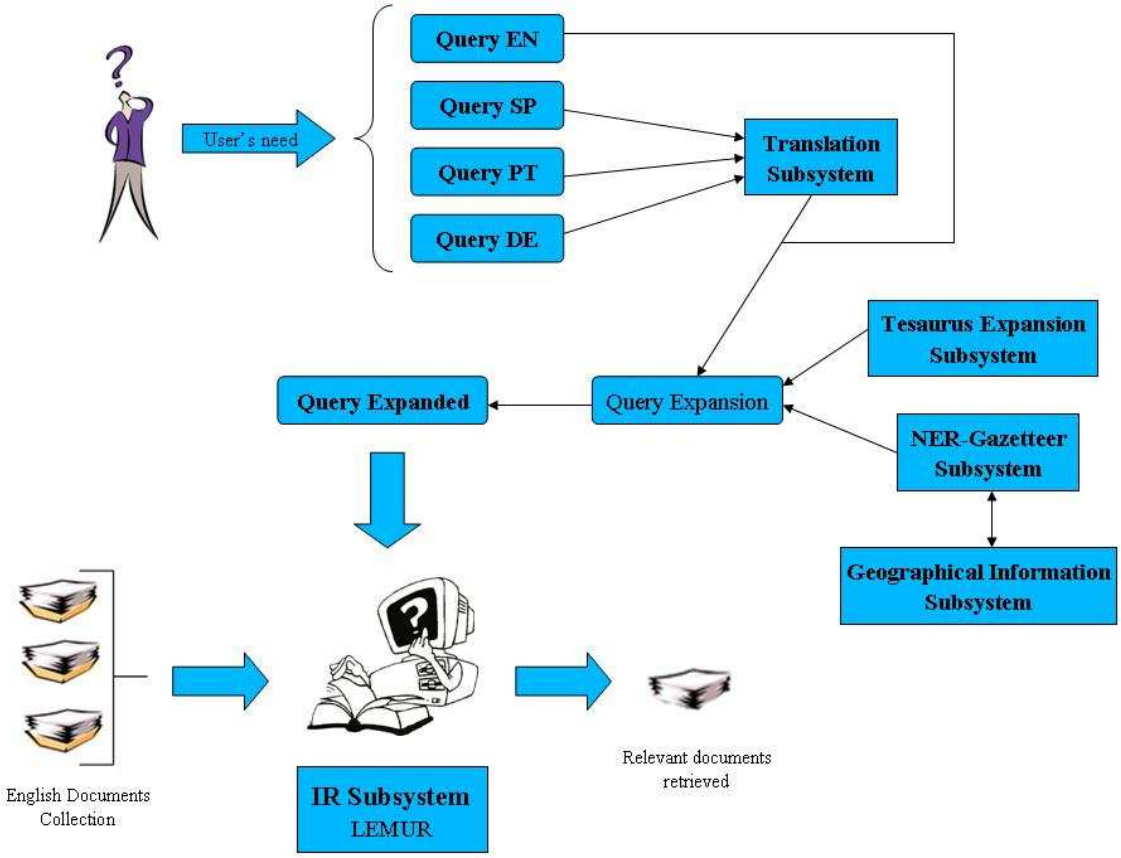

Figure 1: System architecture

2.2 Named Entity Recognition (NER) - Gazetteer Subsystem

The main goal of NER-Gazetteer Subsystem is to detect and recognize the entities in the queries, in order to expand the topics with geographical information. We are only interested in geographical information, so we have used only the locations detected by the NER module. The *location* term includes everything that is *town, city, capital, country* and even *continent*.

As we will see in the next section, the information about locations is loaded previously in the Geographical Information Subsystem, that it is related directly to the NER-Gazetteer Subsystem.

Figure 1 describes the system architecture with this relation. The NER-Gazetteer Subsystem recognizes the entities and provides this information to the Geographical Information module.

We have used the NER system that GATE[3] provides.

The basic operation of NER-Gazetteer module is the following:

- The first step is the preprocessing phase. Each query is preprocessed using a tokenizer, a sentence splitter and a POS tagger. The NER system we have used needs this information in order to improves the named entity detection and recognition.
- The second step is the extraction of the title of each topic and the submission to the NER. The result is saved to another *labelled topic file* with the location entities labelled.
- The last step is the detection of the geographical places, that the NER module have not detected. For this proposal we have use a Gazetteer, included also in GATE. We also

[3] http://gate.ac.uk/

include this information in an expanded topic, using an XML label.

The NER-Gazetteer Subsystem generates some labelled topics, base on the original one, adding the locations.

2.3 Geographical Information Subsystem

The objective of this module is to expand the locations of the topics, using geographical information. We have used *automatic query expansion*[1], a simple expansion that consists in adding terms to the query. We have obtained the geographical information from Geonames gazetteer, a free resource that provides geo-data such as geographical names and postal codes. Its database contains over six million entries for geographical names whereof 2.2 million cities and villages.

Geonames integrates geographical data such as names, altitude, population and others from various sources.

Some examples of queries that receives this module are:

- *Find the capital of a country whose population is greater than X.*
- *Find five cities from a country whose population is greater than X.*
- *Find the country name of a city.*
- *Find the latitude and longitude from a location.*

When a location is recognized by the NER subsystem we look for in the Geographical Information Subsystem.

In addition, it is necessary to consider the spatial relations found in the title ("near to(", "within X miles of(", "north of(", "south of(", etc.). Depending on the spatial relations, the search in the Geographical Information subsystem is more or less restrictive.

We also have to verify if the location is a city, a country or a continent. Depending on the location type, the expansion will become of the following way:

- If the location is a continent, we expand with the capitals of countries that belong to that continent, and with capitals if the population is greater than a number of habitants (one of ours parameters). The expansion is not very large in order to avoid noise in the recovery process.
- If the location is a country, we expand with the five most important cities (with greater population) of that country.
- If the location is a city or capital, first we verified if there is some spatial relation in the topic. If exists we use the latitude and longitude information to find other relevant locations and we expand the topic with them. If there is not some spatial relation, we expand the topic only with the name from the country to which city belongs.

Finally we add the locations given back by the Geographical Information Subsystem to the topic title.

To adjust the parameters of this module, for instance the number of habitants to consider a relevant capital or the number of cities to expand, we have made experiments with the GeoCLEF 2005 framework.

2.4 Thesaurus Expansion Subsystem

A collection of thesauri was generated from the GeoCLEF training corpus. We were looking for words with a very high rate of document co-location. These words will be treated like synonyms and added to the topics.

For that, we generated an inverse file with the entire corpus. The file has a row for each different word of the corpus. Following each word appear all the current word frequencies for each corpus file. These rows can be treated with the standard *TF.IDF* [10] for words comparing test.

We probe this method with the GeoCLEF 2005 corpus and we found that a cosine similarity great than 0.9 between words was the rate that obtain best precision/recall results.

The same procedure was applied to the 2006 corpus. The thesauri collection was generated for all the names of the topics and all the thesauri words were added to its topic. The Figure 2 shows the calculated thesauri for the topics GC033 and GC034. We can see the topic code and the pairs word-similarity.

```
GC033 clisham 1.000 internat 1.000 qbg's 1.000 roinbeabhal 1.000
roineabhal 0.962 roineabh 0.949 anorthosit 0.603 lingerbay 0.585 sport
0.999 competit 1.000 ruhr 1.000 brummer's 0.892 frauenballett 0.892
hyperathlet 0.892 kreisiment's 0.892 ort' 0.892 smokiest 0.892 linke
0.730 hixson's 0.728 itterbeck 0.728 fastman 0.709 ludger 0.564 ort
0.547 urs 0.515 ## world 1.000 championship 1.000 intern 1.000
tournament 1.003 ## ##
GC034 malaria 1.000 plasmodium 0.950 bloland 0.902 heimlich's 0.902
imt 0.902 jauregg 0.902 neurosyphili 0.902 timpone 0.902 trach 0.902
vivax 0.902 wondering! 0.902 heimlich 0.900 audacious 0.856 greentre
0.851 bresler 0.807 lyme 0.563 protozoan 0.521 tropic 0.999 ##
outbreak 1.000 prevent 1.005 vaccin 1.000 ## ##
```

Figure 2: Some examples of a 0.5 similarity thesauri

2.5 Information Retrieval Subsystem

The English collection dataset has been indexed using LEMUR IR system. It is a toolkit[4] that supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or subcollections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

The English collection include a variety of topics and geographical regions form news stories between 1994 and 1995.

Previously the English collection provided for GeoClef 2006 have been preprocessed, using the English *stopwords* list and the Porter *stemmer* [7].

Each topics set, the monolingual expanded and the bilingual translated and also expanded, is run to LEMUR.

One parameter for each experiment is the weighting function, such as Okapi [8] or *TF.IDF*. Another is the use or not of Pseudo-Relevant Feedback (PRF) [9].

3 Experiments and Results

Our baseline experiment is the following:

- We use the original English topics set
- We preprocess each topic (stopper and stemmer)
- Topics without expansion (geographical or thesaurus)
- Information Retrieval without PRF

[4]The toolkit is being developed as part of the Lemur Project, a collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

Experiment	Mean Average Precision	R-Precision
sinaiEnEnExp1 (best result)	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: Summary of results of the monolingual task

3.1 Monolingual tasks

SINAI has participated in monolingual task with five experiments and the official results are shown in Table 1.

Best results were obtained when using our baseline experiment, without adding no type of query expansion (experiment *sinaiEnEnExp1*), only we preprocess each topic with stopper and stemmer and information retrieval process uses Okapi with feedback like weighting function. We considered all tags (*title, description* and *narrative*) in information retrieval process. The experiment *sinaiEnEnExp2* is the same that the previous one but considering only *title* and *description* tags from topics for information retrieval. We use Okapi with feedback like weighting function in all experiments.

In the experiment *sinaiEnEnExp3* we expand only the title of topic with geographical information and considered only *title* and *description* tags from topics for information retrieval with LEMUR system.

In the experiment *sinaiEnEnExp4* we expand the title and the description of topics with thesaurus information, considering only *title* and *description* tags in information retrieval process.

In the experiment *sinaiEnEnExp5* we expand the title and the description from topics with geographical and thesaurus information. We considered only *title* and *description* tags for information retrieval.

3.2 Bilingual tasks

In bilingual task we have participated with a total of five experiments: two experiments for German-English task and three experiments for Spanish-English task. The official results are shown in Table 2.

For German-English task we submit the experiment *sinaiDeEnExp1* in which we have not expanded the title or the description of topics, only we have translated the topic and we preprocess it with stopper and stemmer. The retrieval process uses Okapi with feedback like weighting function too. We considered all tags (*title, description* and *narrative*) for information retrieval process in this experiment. We submit the experiment *sinaiDeEnExp2* for German-English task too. It is the same that previous one but only considering the *title* and *description* tags for information retrieval.

For Spanish-English, in the experiment *sinaiEsEnExp1*, we have not expanded the topics, only we have translated it and we preprocess it with stopper and stemmer. We considered all tags (*title, description* and *narrative*) for information retrieval process in this experiment. We submit the experiment *sinaiEsEnExp2* for Spanish-English task too. It is the same that previous one but only considering the *title* and *description* tags for information retrieval.

Finally, in the experiment *sinaiEsEnExp3*, we expand the title of topics with geographical information, considering only the *title* and *description* tags for information retrieval.

4 Conclusions and Future work

In this paper, we have presented the experiment carried out in our first participation in the GeoCLEF campaign. We have only tried to verify if the topic expansion with geographical and

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

Table 2: Summary of results of the bilingual tasks

thesaurus information increases the effectiveness of the information retrieval process. Evaluation results show that the use of geographical and thesaurus information does not improve the retrieval. But this is the first step for improving the system in the future.

Several reasons exist to explain the worse results obtained with the expansion of topics:

- The NER used sometimes did not work well, because in various topics some entities are recognized and other no. For the future we will try with another NERs.
- In topics, sometimes, appear compound locations like *New England, Middle East, Eastern Bloc,* etc. that are not in Geographical Information Subsystem. Would be interesting to create rules to control this.
- Depending on spatial relation in topic, we could improve the expansion, testing so that cases work better to add more locations or less.

Therefore, we will try to improve the NER-Gazetteer Subsystem and the Thesaurus Expansion Subsystem to obtain one better query expansion.

5 Acknowledgments

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

References

[1] D. Buscaldi, P. Rosso, and E. Sanchis-Arnal. A wordnet-based query expansion method for geographical information retrieval. *Working Notes for the CLEF 2005 Workshop*, 2005.

[2] F. Martínez-Santiago, M.C. Díaz-Galiano, M. García-Vega, M.T. Martín-Valdivia, and L.A. Ureña-López. Sinai on clef 2001: Calculating translation probabilities with semcor. *Advances in Cross-Language Information Retrieval. Editor Carol Peters. Lecture Notes in Computer Science. Springer-Verlag*, pages 185–192, 2001.

[3] F. Martínez-Santiago, M.A. García-Cumbreras, M.C. Díaz-Galiano, and L.A. Ureña-López. Sinai at clef 2004: Using machine translation resources with mixed 2-step rsv merging algorithm. *CLEF 2004 (Cross Language Evaluation Forum)*, 2004.

[4] F. Martínez-Santiago, M.A. García-Cumbreras, Arturo Montejo-Ráez, and L.A. Ureña-López. Sinai at clef 2005: Multi-8 two-years-on and multi-8 merging-only tasks. *Lecture Notes in Computer Science. Springer-Verlag*, 2005.

[5] F. Martínez-Santiago, M.T. Martín-Valdivia, and L.A. Ureña-López. Sinai on clef 2002: Experiments with merging strategies. *Advances in Cross-Language Information Retrieval. Editor Carol Peters. Lecture Notes in Computer Science. Springer-Verlag*, pages 187–197, 2002.

[6] F. Martínez-Santiago, Arturo Montejo-Ráez, M.C. Díaz-Galiano, and L.A. Ureña-López. Sinai at clef 2003: Decompounding and merging. In *CLEF 2003 - Workshop of the Cross-language Evaluation Forum*, Trondheim, Norway, 2003.

[7] M.F. Porter. An algorithm for suffix stripping. In *Program 14*, pages 130–137, 1980.

[8] S.E. Robertson and S.Walker. Okapi-Keenbow at TREC-8. In *Proceedings of the 8th Text Retrieval Conference TREC-8, NIST Special Publication 500-246*, pages 151–162, 1999.

[9] G. Salton and G. Buckley. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences*, 21:288–297, 1990.

[10] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, London, U.K., 1983.