SINAI-GIR System. University of Jaén at GeoCLEF 2008

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

Abstract

This paper describes the third participation of the SINAI research group from University of Jaén 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 have run several 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. On the other hand, the use of query reformulation and geo-expansion does not improve the baseline case either.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Geographic Information Retrieval (GIR), Named Entity Recognition (NER), Spatial Relations, Geo-referencing, Filtering Documents, GeoCLEF

1 Introduction

GeoCLEF is a cross-language Geographic Information Retrieval (GIR) task at the Cross-Language Evaluation Forum (CLEF) campaign since 2005. The aim of GeoCLEF is to evaluate GIR systems. Given a multilingual statement describing a spatial user need (topic), the challenge is to find relevant documents from target collections using queries into several languages [1, 2]. Queries are textual descriptions with three fields (*title*, *description* and *narrative*), including spatial relations and geographic locations such us continents, seas, rivers or regions.

This paper describes the approaches taken by the SINAI[1] research group from the University of Jaén for the main GeoCLEF 2008 subtasks: mono and bilingual retrieval. In 2006 [3], we studied

[1] http://sinai.ujaen.es

the behavior of query expansion using a gazetteer and a thesaurus. In GeoCLEF 2007 [4], we changed the approach and we applied filtering to the documents retrieved by the IR subsystem. This year, our GIR system, called SINAI-GIR, follows a similar architecture of the previous one presented at GeoCLEF 2007 but using new developments related to several techniques such as query reformulation, keywords and hyponyms extraction and even query geo-expansion. We have also implemented new rules for the filtering and re-ranking of documents retrieved.

Next section describes the whole system. Then, in the section 3, each module of the system is explained. Following, results are described and finally, the conclusions about our participation in GeoCLEF 2008 are expounded.

2 System Overview

Our SINAI-GIR system is made up of five main subsystems: *Translator*, *Collection Preprocessing subsystem*, *Query Analyzer*, *Information Retrieval subsystem* and *Validator*. We make use of the Geonames gazetteer[2] as geographic knowledge base for the whole system. As Information Retrieval (IR) index-search engine we have used Lemur[3].

Each translated query is preprocessed and analyzed by the *Query Analyzer*, identifying their geo-entities and spatial relations. This module also applies *query reformulation* based on the query parsing subtask [5], generating several independent queries which will be indexed and searched by means of the IR subsystem. On the other hand, the collection is preprocessed by the *Collection Preprocessing* module and finally the documents recovered by the IR subsystem are filtered and re-ranked by means of the *Validator* subsystem. Figure 1 shows the SINAI-GIR system architecture.

3 Subsystems Description

3.1 Translator

As translation module, we have used SINTRAM (SINai TRAnslation Module), our Machine Translation system which works with different online machine translators and implements several heuristics to combine different translations [6]. This module translates the queries from other languages into English.

3.2 Collection Preprocessing Subsystem

In our architecture we only worked with English documents collection[4] and we have applied a off-line preprocess to this one. In this preprocess we have applied the Porter *stemmer* [7], the English *stop-words* list, the Brill POS tagger [8] and a specific Named Entity Recognizer (NER). The collection preprocessed is indexed later using the IR subsystem.

During the preprocessing, two indexes are generated:

- The **locations index**. This index stores all location entities detected and recognized by the NER in each document of the collection. This year we have used LingPipe[5] as NER in our architecture. All entities classified as a *location* by the NER are checked using the GeoNames gazetteer. This locations index will be used later by the *Validator* in order to filter the documents recovered by the IR subsystem.

[2] http://www.geonames.org/. Geonames is a geographic database which 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.

[3] http://www.lemurproject.org/. Lemur is a open-source IR toolkit which 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.

[4] The collection consists of 169,477 documents composed of stories from the British newspaper *Glasgow Herald* (1995) and the American newspaper *Los Angeles Times* (1994).

[5] http://alias-i.com/lingpipe/. LingPipe is a suite of Java libraries for the linguistic analysis of human language which includes a NER module.

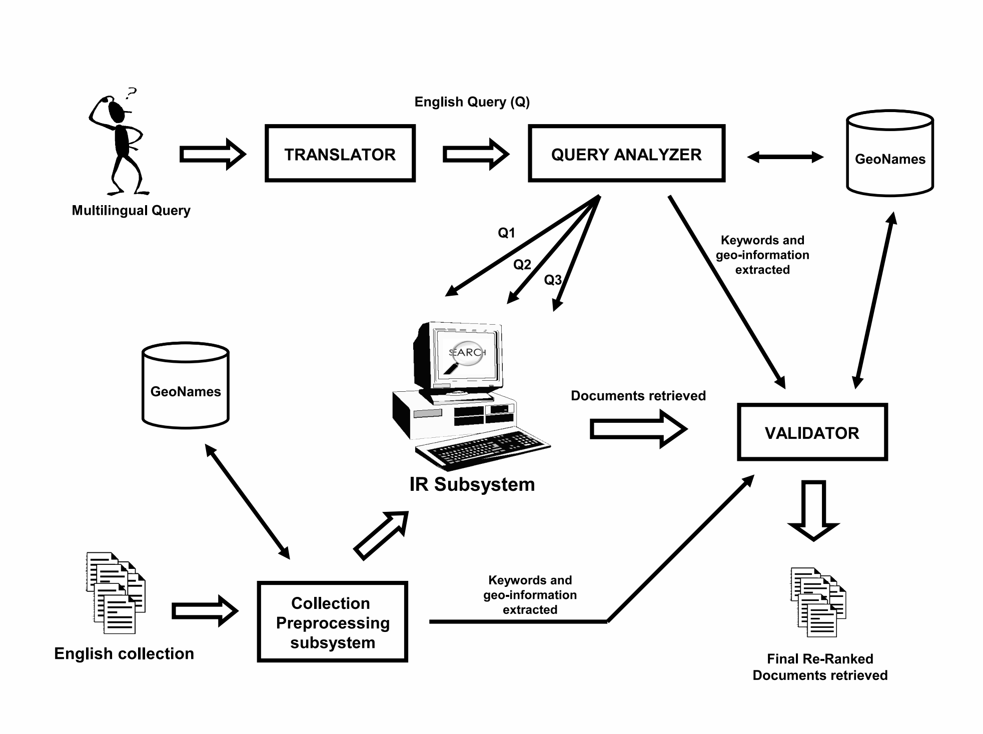
English Query (Q)
Multilingual Query
TRANSLATOR
QUERY ANALYZER
GeoNames
Q1
Q2
Q3
Keywords and geo-information extracted
SEARCH
GeoNames
Documents retrieved
VALIDATOR
IR Subsystem
English collection
Collection Preprocessing subsystem
Keywords and geo-information extracted
Final Re-Ranked Documents retrieved

Figure 1: SINAI-GIR system architecture

- The ***keywords* index**. This index stores the *keywords* of each document of the collection. We have considered as *keyword* the nouns that appearing more than once in the document. We have decided to consider only the nouns because they have more meaning than verbs or adjectives. This *keywords* will be used by the *Validator* in order to re-rank the documents recovered by the IR subsystem.

3.3 Query Analyzer

The Query Analyzer module is responsible for preprocessing of English queries as well as the generation of different query reformulations. This analyzer is also made up of several components:

Preprocessing module. This module removes guidance information from the topics such us "*Relevant documents contain information about*", "*Find documents describing*" or "*To be relevant, documents must describe*", and descriptions about irrelevant documents. The default query (Q_1) is formed by the preprocessed content of the *title*, *description* and *narrative* labels of the topics, applying the Porter *stemmer* [7] and discarding the English *stop-words.*

NER module. The aim of this NER module is to recognize the locations in the queries. As in the *Collection Preprocessing* subsystem, we have used LingPipe. All locations detected are also verified using the GeoNames gazetteer.

Geo-Relation Finder module. This module is used to find the spatial relations in the queries. It is based on manual rules and the entities detected by the NER module. This module makes use of several text files that store all *geo-relations* that can be detect. Some examples of these spatial relations are: *in, near, north of, next to, in or around, in the west of...*

Query Reformulation module. This module parses only the *title* of the query, detecting the three components on which it is usually composed: "*what*", "*geo-relation*" and "*where*". Before the query parsing, we apply a particular *translation* of sentences like "*capital of*

<entity>" or "*<entity>'s capital*", replacing the entire sentence by the corrected location using the Geonames gazetteer. By example, the sentence "*the capital of France*" would be replaced by "*Paris*". This module generates three query reformulations:

- Q_1: it is formed by the preprocessed content of the topic labels, depending on the experiment. Some experiments consider only the content of the *title* and *description* from the topics. Others consider the content of all labels (*title*, *description* and *narrative*).
- Q_2: it is formed only by the concatenation of "*what*" and "*where*" components detected in the *title* of the topic.
- Q_3: it is the same as the previous one, but adding it the expanded locations found in Geonames gazetteer, related to the "*where*" component and depending on the *geo-relation* detected in the *title* of the topic.

Keywords-Hyponyms extractor module. The aim of this module is to detect the *keywords* only in the *title* of the queries. As in the *Collection Preprocessing* subsystem, we have considered as *keywords* only the nouns because they have more importance than verbs or adjectives in the re-ranking process. We have used the Brill POS tagger [8] to detect the nouns. In addition, for each keyword recognized, this module extracts its hyponyms using WordNet[6][9]. These extracted hyponyms also will be used later in the re-ranking process by the *Validator*.

3.4 Information Retrieval Subsystem

The index-search engine that we have used in the experiments is Lemur. It is a open toolkit that supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or sub-collections, and the implementation of retrieval systems based on language models. Lemur supports several weighting functions such as *Okapi* [10], *TF·IDF* and the use of *Pseudo-Relevant Feedback* (PRF) [11]. In the experiments carried out in this paper we have always used the *Okapi with PRF* because it is the weighting function which offers the best results.

The aim of this module is to retrieve the most relevant documents for each query reformulation. As we can see in the previous section, the **Query Analyzer** generates three queries for each topic (Q_1, Q_2 and Q_3). The list of retrieved documents for the Q_1 query is considered as baseline case in our experiments because we do not apply it any filtering or re-ranking process.

3.5 Validator

The aim of the *Validator* process is to filter the lists of documents recovered by the IR subsystem, establishing what of them are valid, depending on the locations and the *geo-relations* detected in the query. Another important function of this process is to establish the final ranking of documents, based on manual rules and predefined weights. These predefined weights allow to assign a score to each document, in addition to the score provided by the IR subsystem.

In order to validate each document recovered, the *Validator* applies different manual rules, making use of geographical data detected in the topic. Some examples of these manual rules are:

- If the entity of the topic is a *country* and its *geo-relation* associated is "*in the north of*", the *Validator* will accept the document if it has any location situated in the north of that country. Using the Geonames gazetteer, the module obtains the maximum and minimum latitudes of all locations that belong to that country. Then, it calculates half of the latitude from the maximal and minimal latitudes to estimate the north of the region.
- If the entity of the topic is a *city* and its *geo-relation* associated is "*near to*", the module will consider that a location is *near to* another one when it is at a distance of less than 50

[6] http://wordnet.princeton.edu/

kilometers around. We have tried several distances in the experiments and the distance of 50 kilometers provided us the best results. In order to measure the distance in kilometers between two locations we have used the simple formula for calculating the distance with geographic coordinates:

$$d = \sqrt{(x^2 + y^2)}$$

where:

$$x = 110.56 * abs(lat2 - lat1)$$

$$y = 84.8 * abs(lon2 - lon1)$$

lat2, *lat1*, *lon2* and *lon1* are the latitudes and longitudes from location 2 and 1 respectively. If the d value is less than 50, the location 1 will be considered *near to* location 2.

- If the entity of the topic is a *continent* or a *country* and its *geo-relation* associated is "*in*", "*of*", "*at*", "*on*", "*from*" or "*along*", the module will accept the document recovered if a location exists in the document that belongs to that continent or country.

In order to re-rank each valid document recovered, the *Validator* makes use of several predefined weights that are added to the score provided by the IR subsystem, depending on the manual rules which the document complies. By example, there is a predefined weight for the first manual rule previously explained (the entity type is *country* and its *geo-relation* associated is "*in the north of*"). If any document complies with that rule, the predefined weight for that rule is added to the IR score of the document. In the experiments carried out in this paper we have tried with several weights, allowing us for an optimal adjustment of the system.

4 Experiments and Results

SINAI has participated in monolingual and bilingual tasks with a total of 15 experiments. In some experiments we have considered the content of all the topics labels (*title*, *description* and *narrative*), identified them as *TDN*, and for others experiments we have considered only the title and description labels, identified them as *TD*.

Our baseline experiment consists of the retrieval of documents more relevant by means of IR subsystem using the Q_1 as query for each topic, without applying any filtering or re-ranking process. This baseline experiment has been applied in the monolingual and bilingual tasks, using *TDN* and *TD* labels from topics.

Other experiments carried out in this paper consist of applying the filtering and re-ranking processes to the different lists of relevant documents retrieved by the IR subsystem. In some experiments, the filtering and re-ranking processes have been applied to the fusion list of the documents recovered by the Q_1, Q_2 and Q_3 queries for each topic. This fusion list is generated by adding the documents from Q_2 and Q_3 lists that are not in Q_1. In other experiments, we have considered the use of *keywords* and *hyponyms* in the re-ranking process.

With respect to the weighting function used in the IR subsystem, in the experiments carried out in this paper we have always used *Okapi with PRF* because after trying several weighting functions, it offers the best results.

4.1 Monolingual task

In monolingual task we have participated with a total of 9 experiments. Three of them are considered as baseline experiments because we have not applied any filtering or re-ranking process to the list of recovered documents. Other experiments combining the *fusion list* and the use of *keywords* and *hyponyms*. The results for monolingual task are shown in Table 1.

Labels	Fusion	Filtering	Keywords	Hyponyms	R-Prec	MAP
TD	no	no	no	no	0.2952	**0.2841**
TDN	no	no	no	no	0.2385	0.2258
TDN	yes	no	no	no	0.2385	0.2250
TD	no	yes	no	no	0.2879	0.2746
TDN	no	yes	no	no	0.2080	0.2119
TDN	yes	yes	no	no	0.1983	0.1960
TD	no	yes	yes	no	0.2828	0.2790
TDN	no	yes	yes	no	0.2176	0.2260
TDN	no	yes	yes	yes	0.2148	0.2221
TDN	yes	yes	yes	no	0.2072	0.2122

Table 1: Summary of results of the monolingual task

Language	Labels	Filtering	R-Prec	MAP
Portuguese	TD	no	0.2365	**0.2183**
Portuguese	TDN	no	0.2028	0.1891
German	TDN	no	0.1127	0.1008
Portuguese	TD	yes	0.2407	0.2166
Portuguese	TDN	yes	0.2014	0.1830
German	TDN	yes	0.1019	0.1161

Table 2: Summary of results of the bilingual task

4.2 Bilingual task

In bilingual task we have participated with a total of 6 experiments. We have used the English collection and the topics in Portuguese and German languages. Three of the experiments are considered as the baseline case because we have not applied any filtering or re-ranking process to them. Other experiments combining the filtering and re-ranking process without the use of *keywords* or *hyponyms*. The results for bilingual task are shown in Table 2.

4.3 Results

The analysis of results based on MAP values shows that the filtering and re-ranking process does not improve the baseline case, because the re-ranking process is not performing well yet. There are valid documents which do not rise enough in the final ranking. Moreover, the number of documents recovered by the IR subsystem normally is around 3000 and the size of the final list returned by the *Validator* is 1000, so some valid documents are left outside for some topics.

However, in some experiments in which we apply filtering, the use of *keywords* improves the result obtained without using *keywords*. Instead, the use of hyponyms does not improve in any case the results.

On the other hand, 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*).

In reference to the results obtained for the bilingual task, we can affirm that the Translator module works better with Portuguese. As in the monolingual experiments, the best results have been achieved without the use of filtering or re-ranking process, although the results using filtering almost equal the baseline experiments.

5 Conclusions

In this paper we have presented the experiments carried out in our third participation in the GeoCLEF track, following the basic architecture used in the 2007 experiments. We have tried to improve the filtering and re-ranking process introduced the previous year, adding new developments related to several techniques such as query reformulation, *keywords* and *hyponyms* extraction and even query geo-expansion. Moreover, we have established predefined weights for each manual rule in *Validator* in order to improve the final score of the valid documents. However, we still get the best result without applying any of these techniques. This is because we have not used an optimal method to raise valid documents in the final ranking, depending on the *geo-information* that recovered documents and topics have in common.

About the new developments employed in the experiments, only the use of *keywords* in the re-ranking process seems to improve the filtering results in some cases. Instead, the use of *hyponyms* does not improve the results. Therefore, the proper use of *keywords* for the re-ranking process could be interesting in the future.

With respect to the experiments in which we have used the *fusion list*, the results obtained indicates that the query reformulation does not seem to work well in this field, although in some topics the Q_2 and Q_3 query types add valid documents to the final list which have not been found by the IR subsystem using the default query (Q_1).

6 Acknowledgments

This work has been supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id_514 granted by the University of Jaén.

References

[1] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, and Paulo Rocha. Geoclef 2006: the clef 2006 cross-language geographic information retrieval track overview. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[2] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. Geoclef 2007: the clef 2007 cross-language geographic information retrieval track overview. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*, 2007.

[3] Manuel García-Vega, Miguel A. García-Cumbreras, L.A. Ureña-López, and José M. Perea-Ortega. GEOUJA System. The first participation of the University of Jaén at GEOCLEF 2006. In *Lecture Notes in Computer Science*, volume 4730 of LNCS Series, pages 913–917. Springer-Verlag, 2007.

[4] José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, and Arturo Montejo-Ráez. GEOUJA System. University of Jaén at GEOCLEF 2007. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*, page 52, 2007.

[5] Zhisheng Li, Chong Wanga, Xing Xie, and Wei-Ying Ma. Query Parsing Task for GeoCLEF 2007 Report. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*, 2007.

[6] Miguel A. García-Cumbreras, L. Alfonso Ureña-López, Fernando Martínez Santiago, and José M. Perea-Ortega. BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. In *Lecture Notes in Computer Science*, volume 4730 of LNCS Series, pages 328–338. Springer-Verlag, 2007.

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

[8] E. Brill. A simple rule-based part-of-speech tagger. In *Proceedings of the third Conference on Applied Natural Language Processing (ANLP'92)*, pages 152–155, Trento, Italy, 1992.

[9] Jason Rennie. Wordnet::querydata: a Perl module for accessing the WordNet database. http://people.csail.mit.edu/∼jrennie/WordNet, 2000.

[10] 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.

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