

The GeoTALP-IR System at GeoCLEF 2005: Experiments Using a QA-Based IR System, Linguistic Analysis, and a Geographical Thesaurus

Daniel Ferrés, Alicia Ageno, and Horacio Rodríguez

TALP Research Center, Software Department
Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08043 Barcelona, Spain
{dferrés, ageno, horacio}@lsi.upc.edu
<http://www.lsi.upc.edu/~nlp>

Abstract. This paper describes GeoTALP-IR system, a Geographical Information Retrieval (GIR) system. The system is described and evaluated in the context of our participation in the CLEF 2005 GeoCLEF Monolingual English task.

The GIR system is based on *Lucene* and uses a modified version of the Passage Retrieval module of the TALP Question Answering (QA) system presented at CLEF 2004 and TREC 2004 QA evaluation tasks. We designed a Keyword Selection algorithm based on a Linguistic and Geographical Analysis of the topics. A Geographical Thesaurus (GT) has been built using a set of publicly available Geographical Gazetteers and a Geographical Ontology. Our experiments show that the use of a Geographical Thesaurus for Geographical Indexing and Retrieval has improved the performance of our GIR system.

1 Introduction

This paper describes GeoTALP-IR, a multilingual Geographical Information Retrieval (GIR) system. The paper focuses on our participation in the CLEF 2005 GeoCLEF Monolingual English task [6].

The GIR system is based on *Lucene*, uses a modified version of the Passage Retrieval module of the TALP Question Answering (QA) system presented at CLEF 2004 [4] and TREC 2004 [5]. We designed a Keyword Selection algorithm based on a Linguistic and Geographical Analysis of the topics. A Geographical Thesaurus (GT) has been build using a set of Geographical Gazetteers and a Geographical Ontology.

In this paper we present the overall architecture of GeoTALP-IR and describe briefly its main components. We also present an evaluation of the system used in the GeoCLEF 2005 evaluation.

1.1 GeoCLEF Task Description

GeoCLEF is a cross-language geographic retrieval task at the CLEF 2005 campaign (consult [6] for more details). The goal of the task is to find as many relevant documents as possible from the document collections, using a topic set. Topics are textual descriptions with the following fields: title, description, narrative, location (e.g. geographical places like continents, regions, countries, cities, etc.) and a geographical operator (e.g. spatial relations like in, near, north of, etc.). See below an example of a topic:

```
<num> GC001 </num>
<orignum> C084 </orignum>
<EN-title> Shark Attacks off Australia and California </EN-title>
<EN-desc> Documents will report any information relating to shark
attacks on humans. </EN-desc>
<EN-narr> Identify instances where a human was attacked by a shark,
including where the attack took place and the circumstances
surrounding the attack. Only documents concerning specific attacks
are relevant; unconfirmed shark attacks or suspected bites are not
relevant. </EN-narr>
<EN-concept> Shark Attacks </EN-concept>
<EN-spatialrelation> near </EN-spatialrelation>
<EN-location> Australia </EN-location>
<EN-location> California </EN-location>
```

2 System Description

2.1 Overview

The system architecture has two phases that are performed sequentially (as shown in Figure 1): Topic Analysis (TA) and Document Retrieval (DR). A collection pre-processing process was carried out in advance.

2.2 Collection Pre-processing

We have used the *Lucene*¹ Information Retrieval (IR) engine to perform the DR task. Before GeoCLEF 2005 we indexed the entire English collections: Glasgow Herald 1995 (GH95) and Los Angeles Times 1994 (LAT94) (i.e. 169,477 documents). We pre-processed the whole collection with linguistic tools (described in the next sub-section) to mark the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process the collection is analyzed with a Geographical Thesaurus (described in the next sub-section). This information was used to build an index (see an example in Figure 2) that contains the following fields for each document:

- **Form Field:** This field stores the original text (word forms) with the Named Entities recognized.

¹ <http://jakarta.apache.org/lucene>

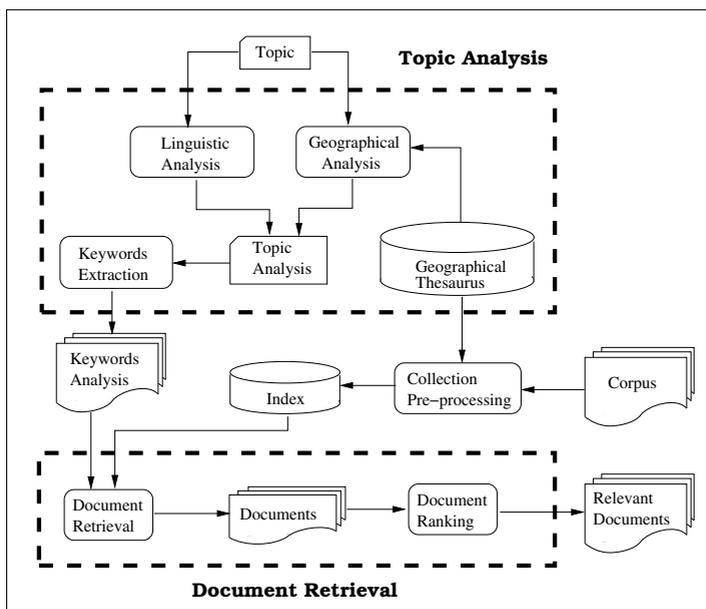


Fig. 1. Architecture of GeoTALP-IR system

Field	Indexed Content
Form	Watson flew off with his wife for a weekend in Barcelona, returned to London on Monday,
Lemma	Watson#NNP#PERSON fly#VBD off#RP with#IN his#PRP\$ wife#NN for#IN a#DT weekend#NN in#IN Barcelona#NNP#LOCATION#city ,#, return#VBD to#TO London#NNP#LOCATION#capital on#IN monday#NNP ,#,
Geo	Europe#Europe#Spain#Cataluña#Barcelona#41.383_2.183 Europe#Europe#United_Kingdom#England#London#51.517_-0.105

Fig. 2. Example of an indexed document

- **Lemma Field:** This part is built using the lemmas of the words, the POS tags, and the results of the Named Entity Recognition and Classification (NERC) module and the Geographical Thesaurus.
- **Geo Field:** It contains all NEs classified as *location* or *organization* that appear in the Geographical Thesaurus. This part has the geographical information about these NE: including geographical coordinates and geographical relations with the corresponding places of its path to the top of the geographical ontology (i.e. a city like "Barcelona" contains its state, country, sub-continent and continent). If a NE is an ambiguous location, all the possible ambiguous places are stored in this field.

2.3 Topic Analysis

The goal of this phase is to extract all the relevant keywords from the topics enriching them as a result of the analysis. These keywords are then used by the Document Retrieval phase. The Topic Analysis phase has three main components: a Linguistic Analysis, a Geographical Analysis and a Keyword Selection algorithm.

Linguistic Analysis. This process extracts lexico-semantic and syntactic information using the following set of Natural Language Processing tools:

- **Morphological components**, a statistical POS tagger (*TnT*) [1] and the WordNet 2.0 [3] lemmatizer are used to obtain POS tags and lemmas. We used the *TnT* pre-defined model trained on the Wall Street Journal corpus.
- **A modified version of the Collins parser**, which performs full parsing and robust detection of verbal predicate arguments [2]. See [5] for more details.
- **A Maximum Entropy based NERC**, a Named Entity Recognizer and Classifier that identifies and classifies NEs in basic categories (person, place, organization and other). This NERC has been trained with the CONLL-2003 shared task English data set [9].
- **Gazetteers**, with the following information: location-nationality relations (e.g. Spain-Spanish) and actor-action relations (e.g. write-writer).

Geographical Analysis. The Geographical Analysis is applied to the Named Entities provided by the location tag (<EN-location>), and the Named Entities from the Title and Description tags that have been classified as *location* or *organization* by the NERC module. This analysis has two main components:

- **Geographical Thesaurus:** this component has been built joining three gazetteers that contain entries with places and their geographical class, coordinates, and other information:
 1. *GEOnet Names Server* (GNS)²: a gazetteer covering worldwide excluding the United States and Antarctica, with 5.3 million entries. Each gazetteer entry contains a geographical name (toponym) and its geographical coordinates (latitude, longitude), language of the geographical name and other features as country, first administrative division,...
 2. *Geographic Names Information System* (GNIS)³, it contains information about physical and cultural geographic features in the United States and its territories. This gazetteer has 2.0 million entries, but we used a subset (39,906) of the most important geographical names.
 3. *GeoWorldMap*⁴ *World Gazetteer*: a gazetteer with approximately 40,594 entries of the most important countries, regions and cities of the world.

² GNS. <http://gnswww.nima.mil/geonames/GNS/index.jsp>

³ GNIS. <http://geonames.usgs.gov/geonames/stategaz>

⁴ Geobytes Inc.: Geoworldmap database containing geographical information and coordinates of cities, regions and countries of the world. <http://www.geobytes.com/>.

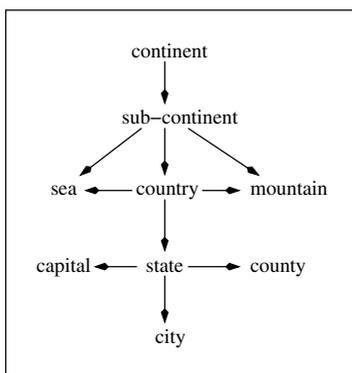


Fig. 3. Geographical ontology

Each one of these gazetteers have a different set of classes. We have mapped this sets to our set of classes (see Figure 3), which includes the most common classes and the most important ones (e.g. country is not common, but is important). The resulting thesaurus contains approximately 3.7 million places with its geographical class. This approach is similar to that used in [7], but they used a limited number of locations (only the 50,000 most important ones).

- **NEC correction filter:** a filter to correct some common errors in the *location-person* and *organization-person* ambiguity classes has been implemented. This filter stores all the NEs classified as *person* in the document; for each one of these NEs it extracts and stores in a hash table all the tokens that compose the NE. Then, for each NE of the document classified as *location* or *organization* it checks whether the NE exists in the document hash. If the NE exists then its class is changed to *person*.

Topic Keywords Selection. We designed an algorithm to extract the most relevant keywords of each topic. These keywords are then passed to the Document Retrieval phase. The algorithm is applied after the Linguistic and Geographical analysis and has the following steps:

1. Initial Filtering. First, all the punctuation symbols and stopwords are removed from the analysis of the title, description and geographical tags.
2. Title Words Extraction. All the words from the title tag are obtained.
3. Description Chunks Filtering. All the Noun Phrase base chunks from the description tag that contain a word with a lemma that appears in one or more words from the title are extracted.
4. Description Words Extraction. The words belonging to the chunks extracted in the previous step and do not have a lemma appearing in the words of the title are extracted.
5. Append Title, Description and Location Words Analysis. The words extracted from the title and description and the geographical tag are appended.

Topic	EN-title	Environmental concerns in and around the Scottish Trossachs
	EN-desc	Find articles about environmental issues and concerns in the Trossachs region of Scotland.
	EN-location	the Scottish Trossachs
Keyword Selection	Title	Environmental concerns Scottish Trossachs
	Stopword Filtering	
	Title Extracted words	Environmental, concerns, Scottish, and Trossachs
	Description Chunks	[environmental issues] [Trossachs region]
	Description Words Extraction	issues and region
Selected Keywords	Environmental#environmental#JJ concerns#concern#NNS issues#issue#NNS region#region#NN scottish#Scottish#NNP#misc#location("Scotland") Trossachs#trossachs#NNP	

Fig. 4. Keyword Selection example

2.4 Document Retrieval

The main function of the Document Retrieval component is to extract relevant documents that are likely to contain the information needed by the user. Document retrieval is performed using the *Lucene* Information Retrieval system. *Lucene* uses the standard tf.idf weighting scheme with the cosine similarity measure, and it allows ranked and boolean queries. The document retrieval algorithm uses a data-driven query relaxation technique: if too few documents are retrieved, the query is relaxed by discarding the keywords with the lowest priority. The reverse happens when too many documents are extracted. Each keyword is assigned a priority using a series of heuristics fairly similar to [8]. For example, a proper noun is assigned a higher priority than a common noun, the adverb is assigned the lowest priority, and stop words are removed.

The main options of the Document Retrieval phase are:

- Query types:
 - Boolean: all the keywords must appear in the documents retrieved. *Lucene* allows boolean queries and returns a score for each retrieved document.
 - Ranked: *Lucene* does ranked queries with tf-idf and cosine similarity.
 - Boolean+Ranked: this mode joins documents retrieved from boolean and ranked queries, giving priority to the documents from the boolean query.
- Geographical Search Mode:
 - Lemma Field: this search mode implies that all the keywords that are Named Entities detected as *location* are searched in the "Lemma" field part of the index.

- Geo Field: this search means that the NEs tagged as *location* and detected as keywords will be searched at the "Geo" index field.
- Geographical Search Policy:
- Strict: this search policy can be enabled when the "Geo" Field search is running, and is used to find a *location* with exactly all this ontological path and coordinates for the following classes: country and region. In example, the form used to search "Australia" in the index is:
Oceania#Oceania#Australia#-25.0_135.0
 - Relaxed: this search policy can also be enabled when the "Geo" field search is running. This mode searches without coordinates. The form used to search "Australia" in the index for this kind of search policy is:
Oceania#Oceania#Australia
In this case, the search is flexible and all the cities and regions of Australia will be returned. An example of a location found with the previous query is:
Oceania#Oceania#Australia#Western_Australia#Perth#-31.966_115.8167

2.5 Document Ranking

This component joins the documents provided by the Document Retrieval phase. If the Query type is *boolean* or *ranked* it returns the first 1000 top documents with their *Lucene* score. In the case of a query mode *boolean+ranked*, it first gives priority to the documents retrieved from the boolean Query and holds their score. The documents provided by the ranked query are added to the list of relevant documents, but their score is then re-scaled using the score of the last boolean document retrieved (the document with lower score of the boolean retrieval). Finally, the first 1000 top documents are selected.

3 Experiments

We designed a set of four experiments that consist in applying different query strategies and tags to an automatic GIR system (see Table 1). Two baseline experiments have been performed: the runs *geotalpIR1* and *geotalpIR2*. These runs differ uniquely in the Query type used: a *boolean+ranked* retrieval in *geotalpIR1* run and only *ranked* retrieval in *geotalpIR2* run. These runs consider the Title and Description tags, and they use the "lemma" index field. The third run (*geotalpIR3*) differs from the previous ones in the use of the Location tag (considering Title, Description and Location) and uses the "Geo" field instead of the "lemma" field. The "Geo" field is used with a Strict Query search policy. This run also performs a *boolean+ranked* retrieval. The fourth run (*geotalpIR4*) is very similar to the third run (*geotalpIR3*), but uses a Relaxed Query search policy.

We can expect a considerable difference between the two first runs and the last ones, because the other ones used an index with geographical knowledge. The fourth run is expected to be better than the third, due to the use of a relaxed search policy, that can increase the recall. On the other hand, we avoided the use

Table 1. Description of the Experiments at GeoCLEF 2005

Run	Run type	Tags	Query Type	Geo. Index	Geo. Search
geotalpIR1	automatic	TD	Boolean+Ranked	Lemma	-
geotalpIR2	automatic	TD	Ranked	Lemma	-
geotalpIR3	automatic	TDL	Boolean+Ranked	Geo	Strict
geotalpIR4	automatic	TDL	Boolean+Ranked	Geo	Relaxed

of the operation tag (e.g. south, in, near,...) because our system is not prepared to deal with this information. Finally, the use of the location tag in the last runs is not so relevant, because our NERC and Geographical Thesaurus are able to detect the place names from the Title and Description tags with high performance.

4 Results

The results of the GeoTalpIR system at the GeoCLEF 2005 Monolingual English task are summarized in Table 2. This table shows the following IR measures for each run: *Average Precision*, *R-Precision*, *Recall*, and the increment over the median of the average precision (0.2063) obtained by all the systems that participated in the GeoCLEF 2005 Monolingual English task.

Table 2. GeoCLEF 2005 results

Run	Tags	AvgP.	R-Prec.	Recall (%)	Recall	Δ AvgP. Diff.(%) over GeoCLEF AvgP.
geotalpIR1	TD	0.1923	0.2249	49.51%	509/1028	-6.78%
geotalpIR2	TD	0.1933	0.2129	49.22%	506/1028	-6.30%
geotalpIR3	TDL	0.2140	0.2377	62.35%	641/1028	+3.73%
geotalpIR4	TDL	0.2231	0.2508	66.83%	687/1028	+8.14%

The results show a substantial difference between the two first runs and the two last ones, specially in the recall measure: 49.51% and 49.22% respectively in the first and second run (*geotalpIR1* and *geotalpIR2*) and 62.35% and 66.38% respectively in the third and fourth run (*geotalpIR3* and *geotalpIR4*). The recall is also improved by the use of Geographical Knowledge and a relaxed policy over the "Geo" Field as it is seen in run four (*geotalpIR4*). Finally, in the last run (*geotalpIR4*) we obtained results about +8.14% better than the median of the average obtained by all runs (0.2063).

5 Conclusions

This is our first attempt to participate in a IR and GIR task. Our approach is based in a QA-based IR system for Document Retrieval and a Linguistic and

Geographical Analysis of the collections and topics. The use of a Geographical Thesaurus has helped to improve the results of our GIR. As a future work, we propose the following improvements to the system: i) analyzing the topics using WordNet, ii) the use of the spatial operator and narrative tags, iii) improving the boolean IR strategy, and iv) the resolution of geographical ambiguity problems.

Acknowledgements

This work has been partially supported by the European Commission (CHIL, IST-2004-506909) and the Spanish Research Department (ALIADO, TIC2002-04447-C02). Daniel Ferrés is supported by a UPC-Recerca grant from Universitat Politècnica de Catalunya (UPC). TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI (the Ministry of Universities, Research and Information Society of the Catalan Government). The authors would like to express their gratitude in particular to Mihai Surdeanu.

References

1. T. Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, Seattle, WA, United States, 2000.
2. M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.
3. Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. pub-MIT, 1998.
4. Daniel Ferrés, Samir Kanaan, Alicia Ageno, Edgar González, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo. The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 557–568. Springer, 2004.
5. Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo. TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints. In *Proceedings of the Text Retrieval Conference (TREC-2004)*, 2005.
6. Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In *Proceedings of the Cross Language Evaluation Forum 2005*, Lecture Notes in Computer Science. Springer, 2006 (in this volume).
7. Manov, D., Kiryakov, A., Popov, B., Bontcheva, K., Maynard, D., Cunningham, H. Experiments with geographic knowledge for information extraction. In *Proceedings of HLT-NAACL Workshop of Analysis of Geographic References*, 2003.
8. D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Gîrju, and V. Rus. LASSO: A tool for surfing the answer net. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, 1999.
9. Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada, 2003.