TALP at GeoCLEF-2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus

Daniel Ferrés and Horacio Rodríguez
TALP Research Center
Software Department
Universitat Politècnica de Catalunya
{*dferres,horacio*}*@lsi.upc.edu*

Abstract

This paper describes our experiments in Geographical Information Retrieval (GIR) in the context of our participation in the GeoCLEF 2006 Monolingual English task. The TALPGeoIR system follows a similar architecture of the GeoTALP-IR system presented at GeoCLEF 2005 [2] with some changes in the Retrieval modes and the Geographical Knowledge Base.

The system has four phases performed sequentially: i) a Keyword Selection algorithm based on a Linguistic and Geographical Analysis of the topics, ii) a Geographical Document Retrieval with Lucene, iii) a Document Retrieval task with the JIRS Passage Retrieval (PR) software, and iv) a Document Ranking phase. A Geographical Thesaurus (GT) has been build using a set of publicly available Geographical Gazetteers and the Alexandria Digital Library (ADL) Feature Type Thesaurus.

In our experiments we have used JIRS, a state-of-the-art PR system for Question Answering (QA), for the GIR task. We also have experimented with an approach using both JIRS and Lucene. In this approach JIRS was used only for Textual Document Retrieval and Lucene was used tor detect the geographically relevant documents. These experiments show that applying only JIRS we obtain better results than combining JIRS and Lucene.

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

Design, Performance, Experimentation

Keywords

Information Retrieval, Passage Retrieval, Geographical Thesaurus, Gazetteers, Feature Type Thesaurus, Named Entity Recognition and Classification

1 Introduction

This paper describes our experiments on Geographical Information Retrieval (GIR) in the context of our participation in the GeoCLEF 2006 Monolingual English task.

GeoCLEF is a cross-language geographic retrieval task at the CLEF 2006 campaign. Like the first GIR task in GeoCLEF 2005 [4], the goal of the GeoCLEF 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.).

Our GIR system is a modified version of the system presented in GeoCLEF 2005 [2] with some changes in the Retrieval modes and the Geographical Knowledge Base. The system has four phases performed sequentially: i) a Keyword Selection algorithm based on a Linguistic and Geographical Analysis of the topics, ii) a Geographical Document Retrieval with Lucene, iii) a Document Retrieval task with the JIRS Passage Retrieval (PR) software, and iv) a Document Ranking phase. A Geographical Thesaurus (GT) has been build using a set of publicly available Geographical Gazetteers and the Alexandria Digital Library (ADL) Feature Type Thesaurus.

In this paper we present the overall architecture of our Geographical IR system and we describe briefly its main components. We also present the experiments, results and conclusions in the context of the GeoCLEF 2006 Monolingual English.

2 System Description

2.1 Overview

The system architecture has two phases that are performed sequentially: Topic Analysis (TA) and Document Retrieval (DR). Previously, a Collection Pre-processing phase has been applied over the textual collections.

2.2 Collection Pre-processing

We pre-processed the entire English collections: Glasgow Herald 1995 (GH95) and Los Angeles Times 1994 (LAT94) (i.e. 169,477 documents) with linguistic tools (described in the next subsection) 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 subsection). This information was used to built two indexes: one with the Geographical information of the documents and another with the Textual and Geographical information of the documents. We have used two Information Retrieval (IR) systems to index: *Lucene*[1] for the Geographical Index and *JIRS* for the Textual and Geographical Index. These indexes are described below:

- **Geographical Index**: this index contains the geographical information of the documents and its Named Entities. The Geographical index contains the following fields for each document:
  - **docid**: this field stores the document identifier.
  - **ftt**: this field indexes the feature type of each geographical name and the Named Entity classes of all the NEs appearing in the document.
  - **geo**: this field indexes the geographical names and the Named Entities of the document. It also stores the geographical information (feature type and geo-ontology path information and coordinates) about the place names. Even if the place is ambiguous all the possible referents are indexed.

[1] http://jakarta.apache.org/lucene

- **Textual and Geographical Index**: this index stores the lemmatized content of the document and adds geographical information (feature type and geo-ontology path information and coordinates) about the Geographical Places of the text. If the geographical place is ambiguous this information is not added to the indexed content.

See below an example of the two indexes:

IR System	Indexed Content	
Lucene	docid	GH950102000000
	ftt	regions@@land_regions@@continents administrative_areas@@political_areas@@countries_1st_order_divisions administrative_areas@@populated_places@@cities administrative_areas@@political_areas@@countries ...
	geo	Europe Asia@@Western_Asia@@Saudi_Arabia@@Hejaz@@24.5_38.5 America@@Northern_America@@United_States@@South_Carolina @@Lodge@@32.9817_-80.952 America@@Northern_America@@United_States@@38.91_-96.19 ...
JIRS	... the role of the wheel in lamatrekking , and where be the good place to air your string vest . pity the crew who accompany him on his travel as sayle of Arabia *countries_1st_order_divisions* **Asia Western_Asia Kuwait Arabia 25.0_45.0** along the Hejaz *countries_1st_order_divisions* **Asia Western_Asia Saudi_Arabia Hejaz 24.5_38.5** railway line from Aleppo *countries_1st_order_divisions* **Asia Middle_East Syria Aleppo 36.0_37.0** in Northern_Syria *countries* **Asia Middle_East Syria 35.0_38.0** to Aqaba *cities* **Asia Western_Asia Jordan Maán Aqaba 29.517_35** in Jordan *countries* **Asia Western_Asia Jordan 31.0_36.0**. as he journey through the searing heat in an age East German ‘ biscuit tin ‘ , his good humour be sorely test ...	

Figure 1: Example of an indexed document with Lucene and JIRS.

2.3 Topic Analysis

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

2.3.1 Linguistic Analysis

This process extracts lexico-semantic and syntactic information using the following set of Natural Language Processing tools: i) **TnT** an statistical POS tagger [1], ii) **WordNet lemmatizer** (version 2.0), iii) **Spear**[2] (a modified version of the Collins parser [3]), and iv) **A Maximum Entropy based NERC** trained with the CONLL-2003 shared task English data set.

2.3.2 Geographical Analysis

The Geographical Analysis is applied to the Named Entities from the Title and Description and Narrative 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 four gazetteers that contain entries with places and their geographical class, coordinates, and other information:

[2]**Spear.** `http://www.lsi.upc.edu/~surdeanu/spear.html`

1. GEOnet Names Server (GNS)[3]: a gazetteer covering worldwide excluding the United States and Antarctica, with 5.3 million entries.
2. Geographic Names Information System (GNIS)[4], contains 2.0 million entries about geographic features of the United States and its territories. We used a subset of 39,906 entries of the most important geographical names.
3. *GeoWorldMap*[5] *World Gazetteer*: a gazetteer with approximately 40,594 entries of the most important countries, regions and cities of the world.
4. *World Gazetteer*[6]: a gazetteer with approximately 171,021 entries of towns, administrative divisions and agglomerations with their features and current population. From this gazetteer we added only the 29,924 cities with more than 5,000 unhabitants.

Each one of these gazetteers have a different set of classes. We have mapped these sets to the ADL Feature Type Thesaurus.

- **Feature Type Thesaurus**. The feature type thesaurus of our Geographical Thesaurus is the ADL Feature Type Thesaurus (ADLFTT). The ADL Feature Type Thesaurus is a hierarchical set of geographical terms used to type named geographic places in English [5]. Both GNIS and GNS gazetteers have been mapped to the ADLFTT, with a resulting set of 575 geographical types. Our GNIS mapping is similar to the one exposed in [5].

2.3.3 Topic Keywords Selection

This algorithm extracts the most relevant keywords of each topic. The algorithm was designed for GeoCLEF 2005 [2]. The algorithm is applied after the Linguistic and Geographical analysis and has the following steps:

1. All the punctuation symbols and stopwords are removed from the analysis of the title, description and narrative tags.
2. All the words from the title tag are obtained.
3. All the Noun Phrase base chunks from the description and narrative tags that contain a word with a lemma that appears in one or more words from the title are extracted
4. The words that pertain to the chunks extracted in the previous step and haven't a lemma appearing in the words of the title are extracted.

Once the keywords are extracted three different keyword sets are created:

- **All**: all the keywords extracted from the topic tags (title, description, and narrative).
- **Geo**: geographical places or geographical types appearing in the topic tags.
- **NotGeo**: all the keywords extracted from the topic tags that are not geographical place names or geographical types.

[3] **GNS**. http://gnswww.nima.mil/geonames/GNS/index.jsp
[4] **GNIS**. http://geonames.usgs.gov/geonames/stategaz
[5] **Geobytes Inc.**: Geoworldmap database containing cities, regions and countries of the world with geographical coordinates. http://www.geobytes.com/.
[6] **World Gazetteer**: http://www.world-gazetteer.com

Topic	EN-title	Wine regions around rivers in Europe
	EN-desc	Documents about wine regions along the banks of European rivers.
	EN-narr	Relevant documents describe a wine region along a major river in European countries. To be relevant the document must name the region and the river
Extracted Keywords Set	Not Geo	wine european
	Geo	Europe#NNP#location#regions@@land_regions@@continents#Europe regions#NN hydrographic_features@@streams@@rivers#NN
	All	wine regions rivers european Europe

Figure 2: Keyword sets sample of Topic 026.

2.4 Geographical Document Retrieval with Lucene

Lucene is used to retrieve geographically relevant documents given a specific Geographical IR query. Lucene uses the standard tf.idf weighting scheme with the cosine similarity measure, and allows ranked and boolean queries. We used boolean queries with a *Relaxed geographical search policy* (see [2] for more details). This search policy allows to retrieve all the documents that have a token that matches totally or partially (a sub-path) the geographical keyword. As an example, the keyword America@@Northern_America@@United_States will retrieve all the U.S. places (e.g. like America@@Northern_America@@United_States@@South_Carolina@@Lodge).

2.5 Document Retrieval using the JIRS Passage Retriever

The JAVA Information Retrieval System (JIRS) software [7] is used to retrieve relevant documents related to a GIR query. JIRS[7] is a PR software specially designed for Question Answering (QA). This system gets passages with a high similarity between the largests n-grams of the question and the ones in the passage. It has 3 modes: simple n-gram model, term weight n-gram model, and distance n-gram model. We used the distance n-gram model. In this model, the weight of a passage is computed using the larger n-gram structure of the question that can be found in the passage itself and the distances among the different n-grams of the question found in the passage.

We used JIRS considering a topic keyword set as a question. We retrieved passages using the n-gram distance model of JIRS with a length of 11 sentences per passage. We obtained a maximum of 100.000 passages per topic. Finally a process selects the relevant documents from the set of retrieved passages. We used two document scoring strategies in order to perform the document selection:

- **Best**: this mode sets as a document score the score of its top-ranked passage from the set of the retrieved passages that belong to this document.
- **Accumulative**: this mode sets as a document score the sum of all the scores of its retrieved passages.

2.6 Document Ranking

This component ranks the documents retrieved by Lucene and JIRS. First, the top-scored documents retrieved by JIRS that appear in the document set retrieved by Lucene are selected. Then, if the set of selected documents is less than 1,000, the top-scored documents of JIRS that not appear in the document set of Lucene are selected with a lower priority than the previous ones. Finally, the first 1,000 top-scored documents are selected. On the other hand, when the system uses only JIRS for retrieval only selects the first 1,000 top-scored documents by JIRS.

[7] **JIRS.** `http://leto.dsic.upv.es:8080/jirs`

3 Experiments

We designed a set of five experiments that consist in applying different IR systems, query keyword sets, and tags to an automatic GIR system (see Table 1). Basically, these experiments can be divided in two groups depending on the retrieval engines used:

- **Only JIRS**. Two baseline experiments have been done in this group: the runs *TALPGeoIRTD1* and *TALPGeoIRTDN1*. These runs differ uniquely in the use of the Narrative tag in the second one. Both runs use one retrieval system, JIRS, and they use all the keywords to perform the query. The experiment *TALPGeoIRTDN3* is similar to the previous experiments but uses a Cumulative scoring strategy to select the documents with JIRS.
- **JIRS & Lucene**. The runs *TALPGeoIRTD2* and *TALPGeoIRTDN2* use JIRS for Textual Document Retrieval and Lucene for Geographical Document Retrieval. Both runs use the *Geo* keywords set for Lucene and the *NotGeo* set for JIRS.

Table 1: Description of the Experiments at GeoCLEF 2006.

Automatic Runs	Tags	IR System	JIRS Keywords	Lucene Keywords	JIRS Score
TALPGeoIRTD1	TD	JIRS	All	-	Best
TALPGeoIRTD2	TD	JIRS & Lucene	NotGeo	Geo	Best
TALPGeoIRTDN1	TDN	JIRS	All	-	Best
TALPGeoIRTDN2	TDN	JIRS & Lucene	NotGeo	Geo	Best
TALPGeoIRTDN3	TDN	JIRS	All	-	Cumulative

In these experiments we can expect to see the difference of these strategies: only JIRS for Geographical and Textual search and JIRS with Lucene for a separated Textual and Geographical Search.

4 Results

The results of the TALPGeoIR system at the GeoCLEF 2006 Monolingual English task are summarized in Table 2. This table has the following IR measures for each run: *Average Precision*, *R-Precision*, and *Recall*.

The results show a substantial difference between the two sets of experiments. The runs that use only JIRS have a better *Average Precision*, *R-Precision*, and *Recall* than the ones that use JIRS and Lucene. The run with the best *Average Precision* is **TALPGeoIRTD1** with 0.1342. The best *Recall* measure is obtained by the run **TALPGeoIRTDN1** with a 68.78% of the relevant documents retrieved. This run has the same configuration of the**TALPGeoIRTD1** run but uses the Narrative tag. Finally, we obtained poor results in comparison with the mean average precision (0.1975) obtained by all the systems that participated in the GeoCLEF 2006 Monolingual English task.

Table 2: TALPGeoIR GeoCLEF 2006 results.

Run	Tags	IR System	AvgP.	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	JIRS	**0.1342**	**0.1370**	60.84%	230/378
TALPGeoIRTD2	TD	JIRS & Lucene	0.0766	0.0884	32.53%	123/378
TALPGeoIRTDN1	TDN	JIRS	0.1179	0.1316	**68.78%**	**260/378**
TALPGeoIRTDN2	TDN	JIRS & Lucene	0.0638	0.0813	47.88%	181/378
TALPGeoIRTDN3	TDN	JIRS	0.0997	0.0985	64.28%	243/378

5 Conclusions

We have applied JIRS, an state-of-the-art PR system for QA, to the GeoCLEF 2006 Monolingual English task. We also have experimented with an approach using both JIRS and Lucene. In this approach JIRS was used only for Textual Document Retrieval and Lucene was used to detect the Geographical relevant documents. The approach with only JIRS was better than the one with JIRS and Lucene combined.

Comparatively with the mean average precision of all runs our Average Precision is a bit low. This fact can be due to several reasons: i) the JIRS PR system may be was not used appropiately or is not suitable for the GIR task, ii) our system is not dealing with geographical ambiguities, iii) our system lacks of query expansion methods, iv) the need of relevance feedback methods, and v) errors in the Topic Analysis phase.

As a future work we propose the following improvements to the system: i) the resolution of geographical ambiguity problems applying toponym resolution algorithms, ii) apply some query expansion methods, iii) study the effect of blind feedback.

Acknowledgments

This work has been partially supported by the European Commission (CHIL, IST-2004-506909). 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 Research Department of the Catalan Government.

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] D. Ferrés, A. Ageno, and H. Rodríguez. The GeoTALP-IR System at GeoCLEF-2005: Experiments Using a QA-based IR System, Linguistic Analysis, and a Geographical Thesaurus. In Peters et al. [6].

[3] 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.

[4] 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 Peters et al. [6].

[5] Linda L. Hill. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 280–290, London, UK, 2000. Springer-Verlag.

[6] C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke., editors. *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers.*, volume 4022 of *Lecture Notes in Computer Science*. Springer, 2006.

[7] José Manuel Gómez Soriano, Manuel Montes y Gómez, Emilio Sanchis Arnal, and Paolo Rosso. A Passage Retrieval System for Multilingual Question Answering. In Václav Matousek, Pavel Mautner, and Tomás Pavelka, editors, *TSD*, volume 3658 of *Lecture Notes in Computer Science*, pages 443–450. Springer, 2005.