TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering

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 2007 Monolingual English task. Our system, called TALPGeoIR, follows a similar architecture of our previous system presented at GeoCLEF 2006 [2] with some changes in the Retrieval modes and the Geographical Knowledge Base.

The system has four phases performed sequentially: i) a Linguistic and Geographical Analysis of the topics, ii) a thematic Document Retrieval search with Terrier, iii) a Geographical Document Retrieval with Geographical Knowledge Bases, iv) a Document Filtering phase.

Our experiments show that Geographical Knowledge Bases can be used to improve the retrieval results of the Terrier state-of-the-art IR system by filtering out non geographically relevant documents.

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

report. 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 2007 Monolingual English task.

GeoCLEF is a cross-language geographic retrieval task at the CLEF 2007 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 2006 [3] with some changes in the Retrieval modes and the Geographical Knowledge Base. The system has four phases performed sequentially: i) a Linguistic and Geographical Analysis of the topics, ii) a thematic Document Retrieval with Terrier, iii) a Geographical Retrieval task with Geographical Knowledge Bases (GKBs), and iv) a Document Filtering phase. In addition, we have developed a toolbox based on 'shape files'[1] for countries, following [8].

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 2007 Monolingual English.

2 System Description

2.1 Overview

The system architecture has four phases that are performed sequentially: Topic Analysis, Textual Retrieval, Geographical Retrieval, and Document Filtering. Previously, a Collection Preprocessing 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 extracted from the documents (and enriched with a GKB) and another with the original textual information. We have used the Terrier Information Retrieval (IR) system to index the Textual Index.

- **Geographical Index**: this index contains the geographical information of the documents: the feature types appearing int the documents and the geographical information associated to each Geographical Named Entity of the document (feature type and geo-ontology path information and coordinates). Even if the place is ambiguous all the possible referents are indexed.
- **Textual Index**: this index stores the lemmatized content of the document without added geographical information

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 two main components: a Linguistic Analysis and a Geographical Analysis.

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) **A Maximum Entropy based NERC** trained with the CONLL-2003 shared task English data set.

[1] `http://www.esri.com`

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 uses a Geographical Knowledge Bases that has three 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:

  1. GEOnet Names Server (GNS)[2]: a gazetteer covering worldwide excluding the United States and Antarctica, with 5.3 million entries.
  2. Geographic Names Information System (GNIS)[3], 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*[4] *World Gazetteer*: a gazetteer with approximately 40,594 entries of the most important countries, regions, and cities of the world.
  4. *World Gazetteer*[5]: 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 ourThe keywords used for

  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].

- **Shape Files Toolbox**.

  [8] propose the use of a publicly available database of 'shape files' for countries. There is a 'shape file' available for each country. Each 'shape file' contains a set of non overlapping regions (represented as polygons), each one consisting of a set of points (X-Y coordinates) representing the 'border' of the area. For most countries the 'shape file' contains only one area but some of them contain more than one, for instance, Italy contains 22 areas (the continental area and several islands).

  In order to cope with 'shape files' we have developped a toolbox (implemented in Prolog) allowing a simple management. The main facilities provided by the toolbox are:

  – Obtaining the border points of a country.
  – Detecting if a point belongs to a country or area.
  – Obtaining a polygon which encodes a certain area of a country using a 9-grid zone division (North, North-West, North-East, West, Central, East, South, South-West, Sout-East).
  – Getting the border points around a point P at a distance D.
  – Getting near points around a point P.

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

2.4 Textual Document Retrieval with Terrier

Terrier is an state-of-the-art Information Retrieval system that includes parameter-free probabilistic retrieval approaches such as: Divergence from Randomness (DFR) models [6], classical TF-IDF weighting, Language Modelling, and Okapi's BM25 probabilistic ranking formula.

This module uses Terrier over a lemmatized index of the document collections and retrieves the relevant documents using the whole content of the tags previously lemmatized.

2.5 Geographical Document Retrieval using Geographical Knowledge Bases

Our Geographical Knowledge Base (described before) is used to retrieve geographically relevant documents given the geographical terms of a Geographical IR query. The GeoKB uses a *Relaxed geographical search policy* (see [2] for more details) over Geographical terms and geographical feature types. 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. In addition, each geographical feature type in the query can be expanded using a set of feature type synonyns and related words that has been manually extracted from the GNIS feature types.

2.6 Document Filtering

This component filters the documents retrieved by Terrier and the Geographical Document Retrieval module. First, the top-scored documents retrieved by Terrier that appear in the document set retrieved by the Geographical Document Retrieval module are selected. Then, if the set of selected documents is less than 1,000, the top-scored documents of Terrier 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 Terrier for retrieval only selects the first 1,000 top-scored documents by Terrier.

2.6.1 Geographical Border Filtering

We developed a new filtering process that using the Shape files toolbox of the GKB allows to create polygons of geographical points that enclose the geographical restriction described by the geographical terms of the topic.

3 Experiments

3.1 Initial Tuning

We performed a set of experiments with the GeoCLEF 2006 topics in order to determine the top performing options for the Terrier IR platform.

The best options were a TF-IDF schema over a lemmatized collection with Porter Stemmer and Query Expansion (docs=10;terms=40). The previous configuration achieved a MAP of 0.3457 in the GeoCLEF 2006. Outperforming the BM25 and the DFR (Divergence From Randomness) schemas with MAPs of 0.3394 and 0.2862.

3.2 Final experiments

For the GeoCLEF 2007 evaluation we designed a set of five experiments that consist in applying, geographical knowledge filtering, Relevance Feedback, and different topic tags to an automatic state-of-the-art IR system (see Table 1). These experiments used the options that gave us the best results with the,

Basically, these experiments can be divided in two groups depending on the retrieval engines used:

- **Only Terrier**. 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 the Terrier IR system without GKBs over a lemmatized collection and applying TFIDF with Porter Stemmer and Query Expansion (docs=10;terms=40) in order to retrieve a max of 10.000 docs per topic.

- **Terrier & GeoKB Filtering**. The runs *TALPGeoIRTD2* and *TALPGeoIRTDN2* use the same Terrier configuration than the previous runs for textual Document Retrieval and a GKB for geographical Document Retrieval. A process of Document Filtering based on a Geographical Document Retrieval re-ranks the textually retrieved docs.
  The experiment *TALPGeoIRTDN3* is similar to the previous experiments but uses Border Filtering and omits Query Expansion with Relevance Feedback. Due to the initial phase of our Border filtering approach, this filtering was only applied to the topics that have a geographical relation that implies "close" or "near" and some regions.

Table 1: Description of the Experiments at GeoCLEF 2007.

Automatic Runs	Tags	IR System	Relevance Feedback	Border Filtering
TALPGeoIRTD1	TD	Terrier	yes	-
TALPGeoIRTD2	TD	Terrier & GeoKB	yes	-
TALPGeoIRTDN1	TDN	Terrier	yes	-
TALPGeoIRTDN2	TDN	Terrier & GeoKB	yes	-
TALPGeoIRTDN3	TDN	Terrier & GeoKB	-	yes

4 Results

The results of the TALPGeoIR system at the GeoCLEF 2007 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 runs that use Terrier and the GeoKB have a better *Average Precision*, *R-Precision* than the ones that use only Terrier. The run with the best *Average Precision* is **TALPGeoIRTD2** with 0.2850. The best *Recall* measure is obtained by the run **TALPGeoIRTDN1** with a 93.23% of the relevant documents retrieved. This run has the same configuration of the**TALPGeoIRTD1** run but uses the Narrative tag. The run **TALPGeoIRTDN3**, that used Border Filtering without Relevance Feedback, shows an slightly improving of the MAP and Recall compared with the results of the other runs that use the Narrative tag: **TALPGeoIRTDN1** and **TALPGeoIRTDN2**.

Table 2: TALPGeoIR results at GeoCLEF 2007.

Run	Tags	IR System	AvgP.	R-Prec.	Recall (%)	Recall
TALPGeoIRTD1	TD	Terrier	0.2711	0.2847	91.23%	593/650
TALPGeoIRTD2	TD	Terrier & GeoKB	**0.2850**	**0.3170**	90.30%	587/650
TALPGeoIRTDN1	TDN	Terrier	0.2625	0.2526	**93.23**%	**606/650**
TALPGeoIRTDN2	TDN	Terrier & GeoKB	0.2754	0.2895	90.46%	588/650
TALPGeoIRTDN3	TDN	Terrier & GeoKB	0.2787	0.2890	92.61%	602/650

5 Conclusions

We used Terrier, a state-of-the-art IR system for QA, to the GeoCLEF 2007 Monolingual English task. We also have experimented with an approach using both Terrier and a Geographical Knowledge Base. In this approach Terrier was used only for textual IR and the GeoKB was used to detect the geographically relevant documents. Our results show that applied GKBs can improve some retrieval results of an state-of-the-art IR system: i) the approach with Terrier and the GeoKB was slightly better in terms of MAP than the one with Terrier alone, ii) the Border Filtering approach applied without Relevance Feedback improved slightly the results in MAP and Recall.

As a future work we propose the following improvements to the system: i) the resolution of geographical ambiguity problems applying toponym resolution algorithms, ii) use Terrier with the Divergence From Randomness algorithm instead of the TFIDF, iii) the improvement and evaluation of the Shape Files toolbox and the Border Filtering algorithm.

Acknowledgments

This work has been supported by the Spanish Research Dept. (TEXT-MESS, TIN2006-15265-C06-05). 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. [7].

[3] Daniel Ferrés and Horacio Rodríguez. TALP at GeoCLEF-2006: Experiments Using JIRS and Lucene with the ADL Feature Type Thesaurus. In Alessandro Nardi, Carol Peters, and Jose Luis Vicedo, editors, *Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006*, September 2006.

[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. [7].

[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] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[7] C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Mller, 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.

[8] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. Geographical information recognition and visualization in texts written in various languages. In *SAC '04: Proceedings of the 2004 ACM symposium on Applied computing*, pages 1051–1058, New York, NY, USA, 2004. ACM Press.