TALP at GeoQuery 2007: Linguistic and Geographical Analysis for Query Parsing

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 on the Geographical Query Parsing pilot-task for English at GeoCLEF 2007. Our system uses some modules of a Geographical Information Retrieval system presented at GeoCLEF 2006 [3] and modified for GeoCLEF 2007. The system uses deep linguistic analysis and Geographical Knowledge to perform the task.

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 the Geographical Query Parsing pilot-task for English at GeoCLEF 2007. The Query Parsing task (GeoQuery) is a pilot task proposed in GEOCLEF 2007. It consists on five subtasks:

- Detect whether the query is geographic or no.
- Extract the WHERE component of the query.
- Extract the GEO-RELATION (from a set of predefined types) if present.
- Extract the WHAT component of the query and classify it as MAP, YELLOW PAGE or INFORMATION types.
- extract the coordinates (LAT-LONG) of the WHERE component. This process involves sometimes a disambiguation task.

As an example, see in Table 1 the information that has to be extracted from the query "Discount Airline Tickets to Brazil".

Field	Content
LOCAL	YES
WHAT	Discount Airline Tickets
WHAT-TYPE	INFORMATION
WHERE	Brazil
GEO-RELATION	TO
LAT-LONG	-10.0_-55.0

In this paper we present the overall architecture of our Geographical Query Parsing system and we describe briefly its main components. We also present the experiments, results and conclusions in the context of the GeoCLEF's 2007 GeoQuery pilot task.

2 System Description

2.1 Overview

The system architecture has two main phases that are performed sequentially: Topic Analysis and Question Classification.

2.2 Topic Analysis

The Topic Analysis phase has two main components: a Linguistic Analysis and a Geographical Analysis.

2.2.1 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, iv) **Spear**[1], a modified version of the Collins parser, which performs full parsing and robust detection of verbal predicate arguments [2] (limited to three predicate arguments: agent, direct object (or theme), and indirect object (benefactive or instrument).

We pre-processed the data-set of 800.000 queries in English from a web search-engine with linguistic tools to obtain the following data structures:

- **Sent**, which provides lexical information for each word: form, lemma, POS tag (Penn-Tree-Bank (PTB) tag-set for English), semantic class of NE, list of EWN synsets and, finally, whenever possible the verbs associated with the actor and the relations between some locations (specially countries) and their gentiles (e.g. nationality).

- **Sint**, composed of two lists, one recording the syntactic constituent structure of the question (basically nominal, prepositional and verbal phrases) and the other collecting the information of dependencies and other relations between these components.

- **Environment.** The environment represents the semantic relations that hold between the different components identified in the question text. These relations are organized into an ontology of about 100 semantic classes and 25 relations (mostly binary) between them. Both classes and relations are related by taxonomic links. The ontology tries to reflect what is needed for an appropriate representation of the semantic environment of the question (and the expected answer). The environment of the question is obtained from *Sint* and *Sent*. A set of about 150 rules was built to perform this task. Refer to [4] for details.

[1] `http://www.lsi.upc.edu/~surdeanu/spear.html`

2.2.2 Geographical Analysis

The Geographical Analysis is applied to the Named Entities from the queries that have been classified as LOCATION or ORGANIZATION by the NERC module. A Geographical Thesaurus is used to extract geographical information about these Name Entities. 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.

A subset of the most important features from this thesaurus has been manually set using 46.132 places (including all kind of geographical features: countries, cities, rivers, states,...). This subset of important features has been used to decide if the query is geographical or not geographical.

2.3 Question Classification

The query classification task is performed through the following steps:

- The query is linguistically preprocessed (as described in the previous subsection) for getting its lexical, syntactic and semantic content. See in Figure 1 the results of the process for the former example. What is relevant in the example is the fine grained classification of 'Brazil' as country, the existence of taxonomic information, both of location type (administrative_areas@@political_areas@@countries) and location content (America@@South_America@@ Brazil), and coordinates (-10.0_-55.0, useful for disambiguating the location and for restricting the search area) and the existence of a shallow syntactic tree consisting on simple tokens and chunks, in this case built by the composition of two chunks, a nominal chunk ('Discount Airline Tickets') and a prepositional one ('to Brazil').

Query: "Discount Airline Tickets to Brazil"
Semantic: [entity(3),mod(3,1),quality(1),mod(3,2),entity(2),i_en_proper_country(5)]
Linguistic: Brazil Brazil NNP LOCATION
Geographical: America@@South_America@@Brazil@@-10.0_-55.0
Feature type: administrative_areas@@political_areas@@countries

Figure 1. Semantic and Geographical Content of GQ-38.

- Over the sint structure, a DCG like grammar consisting of about 30 rules developed manually from the sample of GeoQuery and the set of queries of GeoCLEF 2006, is applied for obtaining the list of topics (each topic represented by its initial and final positions) represented by a triple <geo-relation, initial position, final position>). A set of features (consultive operations over chunks or tokens and predicates on the corresponding sent structures) is used by the grammar. The following features were available:

[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

- **chunk features**: category, inferior, superior, descendents.
- **token features**: num, POS, word form, lemma, NE 1 (general), NE 2 specific.
- **token semantics**: synsets, concrete and generic Named Entity type predicates (Named Entity types include: location, person, organization, date, entity, property, magnitude, unit, cardinal point, and geographical relation.
- **head of the chunk features**: num, POS, word, lemma, first NE, second NE.
- **head of the chunk semantic features**.
- **left corner of the chunk**: num, POS, word form, lemma, NE 1 (general), NE 2 (specific)
- **left corner of the chunk semantics**: WordNet synsets.

See Figure 2 for a sample rule. The rule can be paraphrased as follows: a sentence is composed by two chunks followed by a gap. The first chunk is of type 'npb' or 'np', i.e. it is a nominal phrase, basic or complex, its head cannot be a Named Entity and the limits of the chunk provide the limits of the topic. The second chunk is a 'pp' and it provides the list of locations.

```
parse_sentence(1, DS,CT,CNES) -->
  cc(DS,[(cc,[npb,np]),(hne1,[nil]),(ci,[LI]),(cs,[LS])],[],(1,1)),
  {CT=[(LI,LS)]},
  cc(DS,[(cc,[pp]),(cd,[CD])],[],(1,1)),
  {parse_pp(_,DS,CNES,CD,[])},
  parse_gap(_,DS).
```

Figure 2. Example of DCG rule.

- Finally from the result of step 2 several rule-sets are in charge of extracting: i) LOCAL, ii) WHAT and WHAT-TYPE, iii) WHERE and GEO-RELATION, and iv) LAT-LONG data. So, there are four rule sets with a total of 25 rules. Figure 3 presents an example of a WHAT rule. The rule selects from the list of topics one containing a generic location (e.g. the noun 'city'). In this case the selected topic is assigned to WHAT and the WHAT_TYPE set to 'Map'.

```
classify_question_topic(X,WHAT,'Map'):-
sentence_2(X,(_,CT,_,_)),
CT\==[],
sentence_1(X,S),
member((LC1,LC2),CT),
range(LC1,LC2,R),
member(LC,R),
nth(LC,S,Tk1),
is_ generic_location(Tk1,_),
concatenate_words_pos(X,R,WHAT),!.
```

Figure 3. Example of WHAT classification rule.

3 Experiments and Results

We performed only one experiment for the GeoQuery2007 data set. The experiment consisted in to extracting the requested data for the GeoQuery from a set of 800.000 queries.

The results of the TALP system presented at the GeoCLEF's 2007 GeoQuery Geographical parsing task for English are summarized in Table 1. This table has the following IR measures for each run: *Precision*, *Recall*, and *F1*.

In the evaluation data set, a set of 500 queries had been labeled which are chosen to represent the whole query set (800.000). The submitted results have been manually evaluated using a strict criterion where a correct results should have all <local>, <what>, <what-type> and <where> fields correct (the <lat-long> field was ignored in the evaluation).

Our run achieved the following results: 0.2222 of Precision, 0.249 of Recall, and 0.235 of F1.

Table 1: TALPGeoIR results at GeoQuery 2007.

Team Name	Precision	Recall	F1
TALP	0.222	0.249	0.235

4 Conclusions

This is our first approach to deal with a geographical query parsing task. Our system for the GeoCLEF's 2007 GeoQuery pilot task is based on a deep linguistic and geographical knowledge analysis. Although we need to do further evaluations to compare the system with other ones it seems that our approach could be feasible for the task.

As a future work we propose the following improvements to the system: i) further evaluations of each problem subtask, ii) apply more sophisticated geographical desambiguation algorithms.

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] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing.* PhD thesis, University of Pennsylvania, 1999.

[3] 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 C. Peters, F. C. Gey, J. Gonzalo, G. J.F.Jones, M. Kluck, B. Magnini, H. Mller, and M. de Rijke., editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*. Springer, 2005.

[4] D. Ferrés, S. Kanaan, A. Ageno, E. González, H. Rodríguez, M. Surdeanu, and J. Turmo. The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 557–568. Springer, 2004.