MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR

Sara Lana-Serrano[1], José M. Goñi-Menoyo[1], and José C. González-Cristóbal[1,2]

[1] Universidad Politécnica de Madrid
[2] DAEDALUS - Data, Decisions and Language, S.A.
slana@diatel.upm.es, josemiguel.goni@upm.es,
jgonzalez@dit.upm.es

Abstract. This paper presents the 2005 MIRACLE team's approach to Cross-Language Geographical Retrieval (GeoCLEF). The main goal of the GeoCLEF participation of the MIRACLE team was to test the effect that geographical information retrieval techniques have on information retrieval. The baseline approach is based on the development of named entity recognition and geospatial information retrieval tools and on its combination with linguistic techniques to carry out indexing and retrieval tasks.

1 Introduction

The main objective of the MIRACLE[1] team participation in GeoCLEF task [2] has been to make a first contact with Geographical Information Retrieval systems, focusing most of the effort on the resolution of problems related to the geospatial retrieval: creating multilingual gazetteers, geo-entities recognition, processing spatial queries, document tagging, and document and topic expansion. For information retrieval we have used the set of basic components developed for MIRACLE team [3]: stemming, transformation and filtering.

In the development of the Geographical Information Retrieval system we have used different Information Retrieval models: boolean model for geo-entities recognition, probabilistic model for textual information retrieval, and deterministic model for topic expansion.

2 Geo-entity Recognition

The general task of Named Entity Recognition (NER) involves the identification of proper names in the text and their classification as different types of named entities. The lexical resources that are typically included on a NER system are a lexicon and a grammar. The lexicon stores, using one or more lists, a set of well-known names classified according to their type. The grammar is used for disambiguating the entities that match the lexicon entries on more than one list.

[1] A description of the MIRACLE team can be found in this volume [2].

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 920 – 923, 2006.
© Springer-Verlag Berlin Heidelberg 2006

The geo-entity recognition process that we have developed involves a lexicon consisting of a gazetteer list of geographical resources and several modules for linguistic processing, carrying tasks such as geo-entity identification and tagging.

For lexicon creation we have coalesced two existing gazetteers: the Geographic Names Information System (GNIS) gazetteer of the U.S. Geographic Survey [4] and the Geonet Names Server (GNS) gazetteer of the National Geospatial Intelligence Agency (NGA) [5]. When used together, they meet the main criteria for gazetteer selection we have taken into account: world-wide scope, free availability, open format, location using longitude and latitude coordinates, and homogeneity and high granularity. However, they have some unsuitable properties for our purposes that we have had to improve:

- They use the geographic area as the only criterion to relate resources. We have provided the gazetteers with a flexible structure that allows us to define other types of relationships between resources; for example based on its language (Latin America, Anglo-Saxon countries) or religion (Catholic, Protestant, Islamic,...).
- The top of the hierarchic relationships between resources is the country. It has been necessary to add new features to all the entries to store information on the continent they belong to.
- The entries are in vernacular language. We have selected the most relevant geographic resources (continents, countries, region, counties/provinces and well-known cities) and translated them into English, Spanish and German.

The gazetteer we have been finally working with has 7,323,408 entries. The Lucene [1] information retrieval engine was used for indexing and searching the gazetteers.

The developed named geo-entity identifier involves several stages: text preprocessing by filtering special symbols and punctuation marks, initial delimitation by selecting tokens with a starting uppercase letter, token expansion by searching possible named entities consisting of more than one word, and filtering tokens that do not match exactly any gazetteer entry.

For the geographical entity tagging we have chosen an annotation scheme that allows us to specify the geographical path to the entity. Each one of the elements of this path provides information of its level in the geographical hierarchy (continent, country, region…) as well as a unique identifier that distinguishes it from the rest of the geographical resources of the gazetteer.

3 Topic Expansion

The topic expansion tool developed consists of three functional blocks:

- *Geo-entity Identifier*: identifies geographic entities using the information stored in the gazetteer.
- *Spatial Relation Identifier*: identifies spatial relationships. It can identify the spatial relations defined in a configuration file. Each entry in this file defines both a spatial relationship and its related regular expressions which define patterns for several languages.

- *Expander*: tags and expands the topic in order to identify the spatial relationships and the geo-entities related to them. This block uses a relational database system to compute the points located in a geographic area whose centroid is known.

4 Description of the Experiments

The baseline approach to processing documents and topic queries is made up of the following sequence of steps:

1. *Extraction*: ad-hoc scripts are run on the files that contain particular documents or topic queries collections, to extract the textual data enclosed in XML marks.
2. *Remove accents*: all document words are normalized by eliminating accents in words. This process is done before the stemming one since the gazetteer consists of normalized entity names.
3. *Geo-entity Recognition* or *Topic Expansion*: All document collections and topics are parsed and tagged using the geo-entity recognition tool and the topic expansion tool introduced in the previous section.
4. *Stopwords filter*: all the words known as stop words are eliminated from the document.
5. *Stemming*: the process known as stemming is applied to each one of the words of the document.
6. *Lowercase words*: all document words and tags are normalized by changing all uppercase letters to lowercase.
7. *Indexing*: once all document collections have been processed, they are indexed. We have used two search engines applying them to different experiments: The indexing and retrieval system based on the *trie* data structure developed by the MIRACLE team [3], and the Apache Jakarta Lucene [1] system.
8. *Retrieval*: once all topic queries have been processed and expanded they are fed to the *trie* or Lucene engine for searching the previously built index. In our experiments we have only used OR combinations on the search terms.

This year, we have submitted only runs for monolingual tracks. In addition to the required experiment (identified with the suffix NOR in the run identifier) we have defined four additional experiments. They are differentiated mainly in the search engine used as well as in the topic processing. The experiments whose run identifier has the prefix GC have used the *trie*-based search engine whereas these ones whose run identifier has the prefix LGC have used Lucene system.

The suffix CS and NCS refer to topic processing. For topics processing we have used topic title, topic description and all the geographical tags provided. In the experiments whose run identifier ends in CS, all the topic text has fed the topic expansion process, whereas for the ones that end in NCS we have only used the text from the geographical tag for topic expansion.

Figure 1 shows the results obtained by the experiments. If we analyze the individual topic results, we can mainly derive the following: the topic expansion process in conjunction with OR based searching transforms documents that do not match the geographical criteria of topics into pertinent documents; the use of high granularity gazetters can convert from topics that are assumed precise to ambiguous topics, making the

obtained results considerably worse; and finally, CS experiments provide worse results than NCS experiments since the geo-entity recognition process does not have the capability to distinguish the class of named entities.

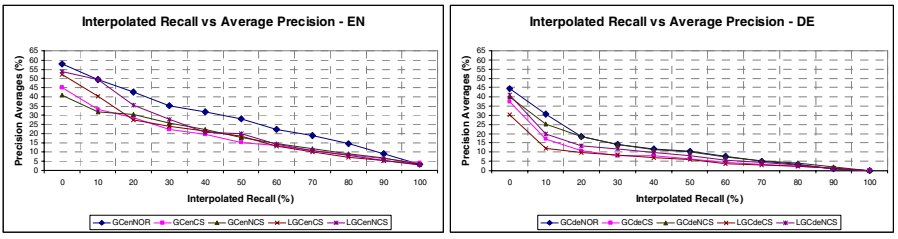

Fig. 1. Results for monolingual English (EN) and German (DE)

5 Conclusions and Future Work

The fundamentals of a geographical information system are the Named Entity Recognition System (NER) in conjunction with the Geographic Information Retrieval (GIR). At this GeoCLEF edition we have tried to attack both aspects of the problem. In order to obtain a solution that approaches better to all the aspects of the problem a great human effort is required.

Future work of the MIRACLE team in this task will be directed to several action lines:

- Improvement of the named entity recognition system adding to it part of speech tagging, classification of the entities and geo-entity disambiguation.
- Incorporation of the improvements obtained by the MIRACLE team, by means of its participation in the ad-hoc track, by using selective or averaging result combination techniques for information retrieval.

References

1. Apache Lucene project. On line http://lucene.apache.org [Visited 2005/10/10].
2. Gey, Frederic; Larson, Ray; Sanderson, Mark; Joho, Hideo; Clough, Paul; and Petras, Vivien. GeoCLEF: The CLEF 2005 Cross-Language Geographical Information Retrieval Track Overview. Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 (in this volume).
3. Goñi-Menoyo, José M.; González, José C.; and Villena-Román, J.. MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 (in this volume).
4. U.S. Geological Survey. On line http://www.usgs.gov [Visited 2005/10/10].
5. U.S. National Geospatial Intelligence Agency. On line http://www.nga.mil [Visited 2005/10/10].