MIRACLE's 2005 Approach to Geographical Information Retrieval

Sara Lana-Serrano[1], José M. Goñi-Menoyo[1]
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's team 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 cause to 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 perform indexing and retrieval tasks.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval ; H.3.4 Systems and Software. E.1 [Data Structures]; E.2 [Data Storage Representations]. H.2 [Database Management]: H.2.5 Heterogeneous Databases; H.2.8 Database Applications - *Spatial databases and GIS*.

Keywords

Geographical IR, geographic entity recognition, spatial retrieval, gazetteer, linguistic engineering, information retrieval, *trie* indexing.

Introduction

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a company founded in 1998 as a spin-off of two of these groups. DAEDALUS is a leading company in linguistic technologies in Spain and is the coordinator of the MIRACLE team. This is the third participation in CLEF, after years 2003 and 2004 [4], [5], [7], [8], [9], [10],[11], [17], [18]. As well as GeoCLEF tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and bilingual, monolingual and cross lingual tracks.

In GeoCLEF task the objective is to evaluate Geographical Information Retrieval (GIR) system involving both spatial and multilingual aspects. The main challenges in the development of a system of these characteristics are the side aspects of the main problem of geographical information retrieval in a multilingual environment (translating locations, ambiguity of geo-references, finding/creating a multilingual gazetteer…) and the inherent ones to the information retrieval (stemming, transformation, filtering, generation of n-grams, relevance feedback, indexing…).

The main objective of the MIRACLE team participation in GeoCLEF task has been to have 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 [5]: stemming, transformation (transliteration, elimination of diacritics and conversion to lowercase) and filtering (elimination of stop and frequent words). A more in-depth description of the MIRACLE toolbox used for pre-processing and indexing the document collections required for this track can be found in the paper "MIRACLE's 2005 Approach to Monolingual Information Retrieval" that can be found in this on-line documentation.

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.

For this year, we have submitted runs for the following tracks:

a) Monolingual English.

b) Monolingual German.

1 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 (persons, organizations, locations). The lexical resources that are typically included in 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 in more than one list.

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.

Gazetteer creation

A gazetteer is an index or geographical directory consists of geo-gazetteer entries that define natural and cultural features with one o more names in one or more languages, sets of coordinates, feature designations, hierarchical relationships and complementary information.

For lexicon creation we have coalesced two existing gazetteers: the Geographic Names Information System (GNIS) gazetteer of the U.S. Geographic Survey [15] and the Geonet Names Server (GNS) gazetteer of the National Geospatial Intelligence Agency (NGA) [16]. 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, countries Anglo-Saxon) 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 about 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 more popular cities) and translated them into English, Spanish and German languages. For this task, manual translating has been combined with the Systran software [13].

The gazetteer we have been finally working with has 7,323,408 entries, each one characterized by several features, such as unique identifier, continent, country, longitude, latitude, name, etc.

The information retrieval engine used for indexing and searching the gazetteers has been Lucene [2]. Lucene is a freely available open-source from the Apache Jakarta project. Lucene supports a Boolean query language, performs ranked retrieval using the standard *tf.idf* weighting scheme with the cosine similarity measure and allows content tagging by treating documents as collections of fields.

Named Geo-entity Identification

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.

Named Entity Tagging

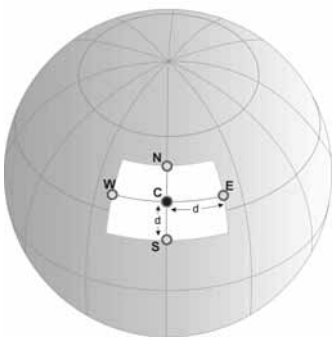
For the geographical entities tagging we have chosen an annotation scheme that allows us to specify the geographical path to the entity. Each one of the elements of this path provides information of its level in the geographical hierarchy (continent, country, region…) as well as a unique identifier that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

2 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.

  The expansion made by the algorithm is determined by the type of geographic resource (continent, country, region, county, city…) and the associated spatial relation. Table 1 shows the different space relations supported by the algorithm and the expansion conducted for each one from them.



All the expansions are based on determining the existing geographical resources in a space region delimited by, at least, three of the following points: N, S, E and W, where:

- C is the centroid of the geographic resource.
- N is the point locates *d* km north of C.
- S is the point locates *d* km south of C.
- E is the point locates *d* km east of C.
- W is the point locates *d* km west of C.
- *d* depend on the resource and spatial relation.

3 Description of the experiments

The baseline approach to processing documents and topic queries is composed 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. We have used HEADLINE and TEXT marks for document collections and the TITLE, DESC, CONCEPT, SPATIALRELATION and LOCATION marks for topics. The contents inside these marks were concatenated to feed the followings steps.
2. **Remove accents:** all documents words are normalized by eliminating accents in words. In spite of this process provides better results running it before the stemming step, we have had to do in this order because our 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. **Lowercase words:** all document words and tags are normalized by changing all uppercase letters to lowercase.
5. **Stopwords filter:** all the words known as stop words are eliminated from the document.
6. **Stemming:** the process known as stemming is applied to each one of the words of the document.
7. **Indexing:** once all document collections have been processed, they are indexed. For this GeoCLEF edition we have used the two following search engines applying them to different experiments:
   - Indexing and retrieval system based on the *trie* [1] data structure developed by MIRACLE team during the two last years [6].
   - Lucene system from the Apache Jakarta project.

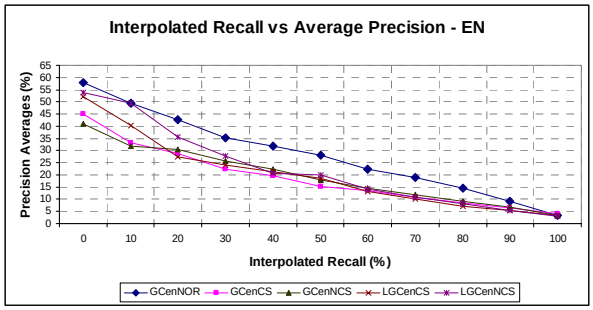
8. **Retrieval:** once all topic queries have been processed and expanded they are fed to the *trie* or Lucene engine for searching the previously built index. In our experiments we have only used OR combinations on the search terms.

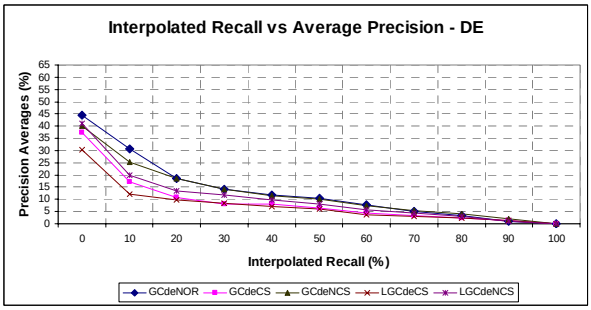
For running most of the previous steps, we have used the set of basic components developed by MIRACLE team [5] adapting them when needed. We have used Porter [12] stemmers and some resources from Neuchatel [14].

For 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 end in CS, all the topic text has fed the topic expansion process, whereas for the ones that end in NCS we have used only the text from the geographical tag for topic expansion.

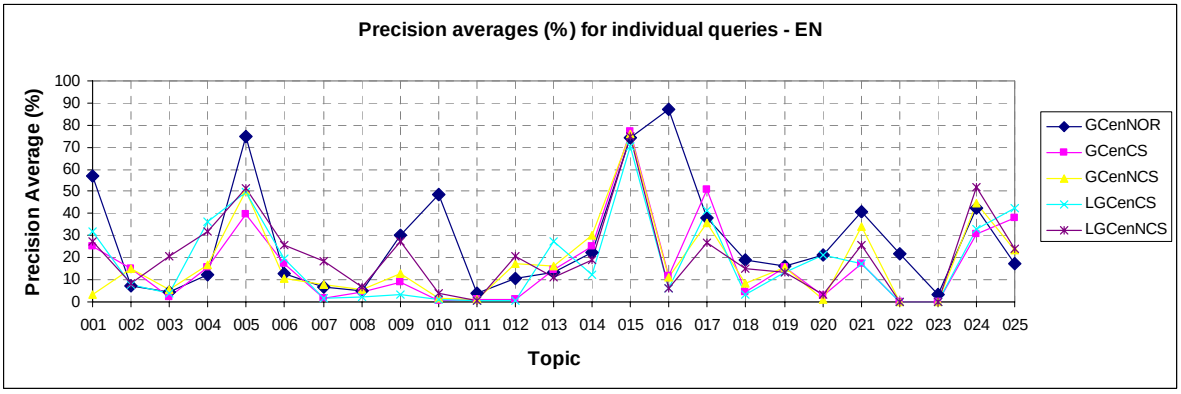
The following figures show the results obtained by the experiments in monolingual English (EN) and monolingual German (DE) tasks.

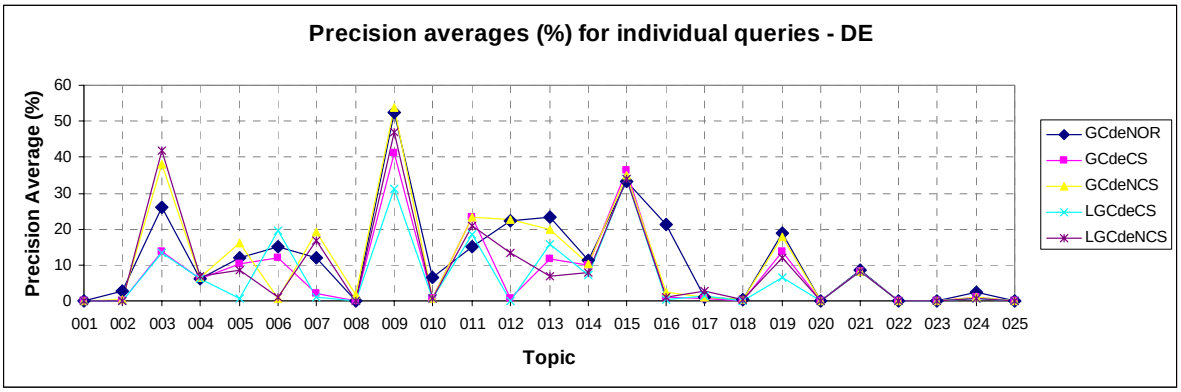




If we analyze the individual topic results, we observe that the topic expansion improves slightly the precision results for some topics, but it gets worse for others. For a topic such as '…*rice imports in Japan*…', the topic expansion process, in conjunction with OR based searching, transforms documents with any Japanese resources into pertinent documents. In other topics, such as topic number 016, in which an ambiguous query (…*oil prospecting in Siberia*…) meets a high granularity in the gazetteer, the topic expansion produces considerably worse results (our gazetteer stores 47 different resources named exactly *Siberia*).

We can assert that CS experiments provide worse results than NCS experiments. This fact can be explained since the geo-entity recognition process do not have the capability to distinguish the class of named entities outcoming noise.





4 Conclusions

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. For this reason we have obtained only one first approach that will be necessary improved.

Nevertheless, in spite of the drawbacks of our solution, we consider that the set of topics selected for the experiments are not very suitable to evaluate the kindness of such approach, due to the small number of pertinent documents. This fact has had a negative impact on the evaluation of the performance of the module of geospatial relationships processing.

5 Future work

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 bilingual, monolingual and cross lingual tracks, by using selective or averaging result combination techniques for information retrieval.

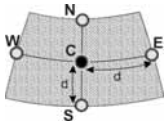
Acknowledgements

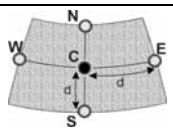
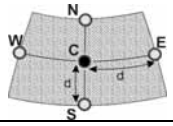
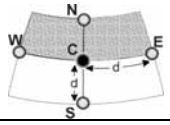
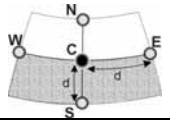
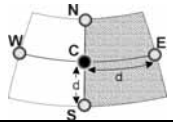
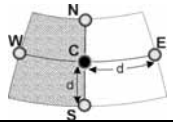
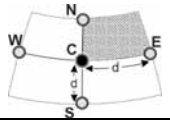
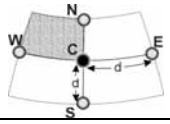
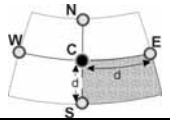
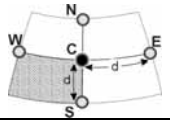
This work has been partially supported by the Spanish R+D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01.

Special mention to our colleagues of the MIRACLE team should be done (in alphabetical order): Ana María García-Serrano, Ana González-Ledesma, José Mª Guirao-Miras, José Luis Martínez-Fernández, Paloma Martínez-Fernández, Ángel Martínez-González, Antonio Moreno-Sandoval and César de Pablo-Sánchez.

Appendix: Tables and figures

Table 1: Topic expansion

Spatial Relation	Example English	Expansion
NORMAL	Madrid	Resource tag.
IN	in Madrid	Resource tag.
NEAR	near to Madrid near Madrid next to Madrid next Madrid	Expansion if not administrative region. 

Spatial Relation	Example English	Expansion
IN_NEAR	in or around Madrid in and around Madrid	Resource tag if continent, country, county, province or borough and expansion if otherwise. 
DISTANCE	within *d* mile/s of Madrid within *d* kilometer/s of Madrid	Expansion if not administrative region. 
NORTH	north of Madrid	Expansion if not administrative region. 
SOUTH	south of Madrid	Expansion if not administrative region. 
EAST	east of Madrid	Expansion if not administrative region. 
WEST	west of Madrid	Expansion if not administrative region. 
NORTH_EAST	northeastern of Madrid northeast of Madrid	Expansion if not administrative region. 
NORTH_WEST	northwestern of Madrid northwest of Madrid	Expansion if not administrative region. 
SOUTH_EAST	southeastern of Madrid southeast of Madrid	Expansion if not administrative region. 
SOUTH_WEST	southwestern of Madrid southwest of Madrid	Expansion if not administrative region. 

References

[1] Aoe, Jun-Ichi; Morimoto, Katsushi; Sato, Takashi. An Efficient Implementation of Trie Structures. Software Practice and Experience 22(9): 695-721, 1992.

[2] Apache Lucene project. On line http://lucene.apache.org [Visited 17/08/2005].

[3] Automatic Trans SL, Spain. Automatic translation server. On line http://www.automatictrans.es [Visited 28/07/2005].

[4] Goñi-Menoyo, José M; González, José C.; Martínez-Fernández, José L.; and Villena, J. MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005 (to appear).

[5] Goñi-Menoyo, José M.; González, José C.; Martínez-Fernández, José L.; Villena-Román, Julio; García-Serrano, Ana; Martínez-Fernández, Paloma; de Pablo-Sánchez, César; and Alonso-Sánchez, Javier. MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. Working Notes for the CLEF 2004 Workshop (Carol Peters and Francesca Borri, Eds.), pp. 141-150. Bath, United Kingdom, 2004.

[6] Goñi-Menoyo, José Miguel; González-Cristóbal, José Carlos and Fombella-Mourelle, Jorge. An optimised trie index for natural language processing lexicons. MIRACLE Technical Report. Universidad Politécnica de Madrid, 2004.

[7] Martínez-Fernández, José L.; García-Serrano, Ana; Villena, J. and Méndez-Sáez, V.; MIRACLE approach to ImageCLEF 2004: merging textual and content-based Image Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

[8] Martínez, José L.; Villena, Julio; Fombella, Jorge; G. Serrano, Ana; Martínez, Paloma; Goñi, José M.; and González, José C. MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 210-219. Springer, 2004.

[9] Martínez, J.L.; Villena-Román, J.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.): Evaluation of MIRACLE approach results for CLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[10] de Pablo, C.; Martínez-Fernández, J. L.; Martínez, P.; Villena, J.; García-Serrano, A. M.; Goñi, J. M.; and González, J. C. *miraQA*: Initial experiments in Question Answering. Working Notes for the CLEF 2004 Workshop, pp. 405-411 (Carol Peters and Francesca Borri, Eds.), pgs. 371-376. Bath, United Kingdom, 2004.

[11] de Pablo, C.; Martínez-Fernández, J. L.; Martínez, P.; Villena, J.; García-Serrano, A. M.; Goñi, J. M.; and González, J. C. *miraQA*: Initial experiments in Question Answering. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

[12] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 13/07/2005].

[13] SYSTRAN Software Inc., USA. SYSTRAN 5.0 translation resources. On line http://www.systransoft.com [Visited 13/07/2005].

[14] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 13/07/2005].

[15] U.S. Geological Survey. On line http://www.usgs.gov [Visited 17/08/2005].

[16] U.S. National Geospatial Intelligence Agency. On line http://www.nga.mil [Visited 17/08/2005].

[17] Villena, Julio; Martínez, José L.; Fombella, Jorge; G. Serrano, Ana; Ruiz, Alberto; Martínez, Paloma; Goñi, José M.; and González, José C. Image Retrieval: The MIRACLE Approach. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 621-630. Springer, 2004.

[18] Villena-Román, J.; Martínez, J.L.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.); MIRACLE results for ImageCLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.