Report of MIRACLE team for Geographical IR in CLEF 2006

Sara Lana-Serrano[1,2], 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

The main objective of the designed experiments is testing the effects of geographical information retrieval from documents that contain geographical tags. In the designed experiments we try to isolate geographical retrieval from textual retrieval replacing all geo-entity textual references from topics with associated tags and splitting the retrieval process in two phases: textual retrieval from the textual part of the topic without geo-entity references and geographical retrieval from the tagged text generated by the topic tagger. Textual and geographical results are combined applying different techniques: union, intersection, difference, and external join based.

Our geographic information retrieval system consists of a set of basics components organized in two categories: (i) linguistic tools oriented to textual analysis and retrieval and (ii) resources and tools oriented to geographical analysis. These tools are combined to carry out the different phases of the system: (i) documents and topics analysis, (ii) relevant documents retrieval and (iii) result combination.

If we compare the results achieved to the last campaign's results, we can assert that mean average precision gets worse when the textual geo-entity references are replaced with geographical tags. Part of this worsening is due to our experiments return cero pertinent documents if no documents satisfy de geographical sub-query. But if we only analyze the results of queries that satisfied both textual and geographical terms, we observe that the designed experiments recover pertinent documents quickly, improving R-Precision values.

We conclude that the developed geographical information retrieval system is very sensible to textual geo-reference and therefore it is necessary to improve the name entity recognition module.

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; H.3.7 Digital libraries. **H.2 [Database Management]**: H.2.5 Heterogeneous Databases; H.2.8 Database Applications - *Spatial databases and GIS*. **E.1 [Data Structures]**: *trees*; **E.2 [Data Storage Representations]**.

Keywords

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

1. 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 our fourth participation in CLEF, after years 2003, 2004, and 2005. As well as GeoCLEF task, the team has participated in the ImageCLEF, Q&A, and bilingual, monolingual and robust multilingual tracks.

The main objective of the designed experiments is testing the effects of geographical information retrieval from documents that contain geographical tags. In the designed experiments try to isolate geographical retrieval from textual retrieval replacing all geo-entity textual references from topics with associated tags and splitting the retrieval process in two phases: textual retrieval from the textual part of the topic without geo-entity references and geographical retrieval from the tagged text generated by the topic tagger. Textual and geographical results are combined applying different techniques: union, intersection, difference and external join based.

We have submitted runs for the following tracks:

- Monolingual English.
- Monolingual German.
- Monolingual Spanish.

This paper is organized as follow: first of all, we briefly present the main components of our geographical information retrieval system, next we describe the experimental runs and results, and finally, we expound the conclusions and future works for GeoCLEF.

2. System Description

Our geographic information retrieval system consists of a set of basics components organized in two categories:

- Linguistic tools oriented to textual analysis and retrieval.
- Resources and tools oriented to geographical analysis.

In this section we describe the main geographical tools of our approach. They make up the Named Entity Recognition (NER) module and the tagging module.

2.1. Gazetteer

The geo-entity recognition system developed involves a lexicon consisted 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 [10] and the Geonet Names Server (GNS) gazetteer of the National Geospatial Intelligence Agency (NGA) [11].

Each one of them uses its own scheme of conceptualization and characterization of the resources. We have defined our own geographical ontology and we had completed and joined the referred gazetteers to obtain a more accurate and flexible gazetteer. The defined ontology allows locating resources based on its geographical areas as well as on other types of relationships like its language (Latin America, countries Anglo-Saxon) or religion (catholic, protestant, Islamic ...).

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, county/region, 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 manages structured information treating documents as collections of fields.

2.2. Named Geo-entity Identifier

The named geo-entity identify process 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 eliminating tokens that do not match exactly any gazetteer entry.

The identifier associates each identified named geo-entity with a list of matched resources of the gazetteer.

2.3. Named Entity Tagger

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 an unique identifier that distinguishes it from the rest of geographical resources of the gazetteer.

3. Description of the MIRACLE experiments

The designed experiments consist of a set of tasks that must be executed sequentially:

- Documents and topics analysis.
- Relevant documents retrieval.
- Result combination.

3.1. Document analysis

The baseline approach to processing documents is composed of the following sequence of steps:

1. **Extraction:** ad-hoc scripts are run on the files that contain particular documents collections to extract the textual data enclosed in XML marks. We have used HEADLINE and TEXT marks. The contents inside these marks were concatenated to feed the followings steps.
2. **Remove accents:** all document 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:** all document collections and topics are parsed and tagged using the geo-entity recognition and tagging tool introduced in the previous section.
4. **Tokenization:** this process extracts basic text components, detecting and isolating punctuation symbols. Some basic entities are also treated, such as numbers, initials, abbreviations, and years. For now, we do not treat compounds, proper nouns, acronyms or other entities. The outcomes of this process are only single words and years that appear as numbers in the text (e.g. 1995, 2004, etc.)
5. **Lowercase words:** all document words and tags are normalized by changing all uppercase letters to lowercase.
6. **Filtering:** all words recognized as *stopwords* are filtered out. *Stopwords* in the target languages were initially obtained from [9], but were extended using several other sources and our own knowledge and resources.
7. **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. We have used standard stemmers from Porter [7].
8. **Indexing:** a *trie* [1] based indexing and retrieval engine developed by MIRACLE [4] has been used for all experiments.

3.2. Topic analysis

This task is compounded of the following phases:

Conversion to structured topic: this phase parses the original topic in a structured XML topic consisting of two main entities. The first one (text-query) only contains the associated textual query with the topic, without textual geo-entity references. The second one (geo-query) is a complex XML entity that describes the geo-entity and geo-spatial references extracted from the topic. This second XML entity tries to disambiguate the geo-entities using only information contained in the original topic.

Geo-query expansion: this phase parses the geo-query XML entity and returns a textual document with geographical tags related to the query. The expansion tool developed for this purpose consists of three functional blocks:

- **Geo-entity Identifier**: identifies geographic entities using the information stored in the gazetteer.
- **Spatial Relation Identifier**: identifies and qualifies spatial relationships supported by a regular expression based system.
- **Expander**: tags and expands the topic according to the identified 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 [5][6].

Text-query processing: the goal of this phase is to apply to the textual query similar linguistic processing that has been applied to the collection documents: punctuation marks and accents elimination, conversion to lowercase letters, filtered of stopwords, and stemming.

3.3. Relevant documents retrieval

When all the documents and topics have been processed, they are fed to an ad-hoc front-end of the retrieval *trie* engine to search the built document collection index. Only OR combinations of the search terms were used. The retrieval model used is the well-known Robertson's Okapi BM-25 [8] formula for the probabilistic retrieval model, without relevance feedback.

3.4. Results combination

The textual and geographical results from topics can be combined applying different techniques: union (OR), intersection (AND), difference (AND NOT), and external join (LEFT JOIN) based. Each of these techniques reranks output results computing new relevance measure value from input results.

4. Basic experiments

For this campaign we have designed several basic experiments where the documents for indexing and the topic queries for retrieval are processed using the combination of the steps described in the previous section. They are differentiated mainly by the topic processing as well as by the results combining.

We have used in the run identifier the following letters to denote the fields used in the text-query (t) and geo-query (g) processing:

- **N:** title, and description fields.
- **A:** title, description, and narrative fields.

We have used the following letters to describe the combining partial results:

- **O:** OR-based combination.
- **A:** AND-based combination.
- **L:** LEFT-JOIN based combination.

Next table shows combinations applying on the submitted runs.

Run Identifier	Text-query (t)	Geo-query (g)	Combination
AA	Title, description, narrative	Title, description, narrative	t AND g
AO	Title, description, narrative	Title, description, narrative	t OR g
AtLg	Title, description, narrative	Title, description, narrative	t LEFT JOIN g
(mandatory run) NA	Title, description	Title, description	t AND g
NtLg	Title, description	Title, description, narrative	t LEFT JOIN g

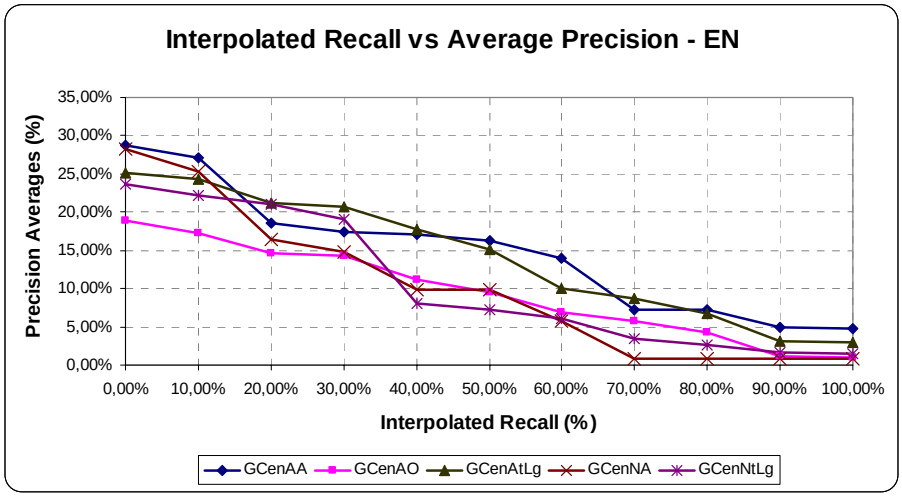
5. MIRACLE results for CLEF 2006

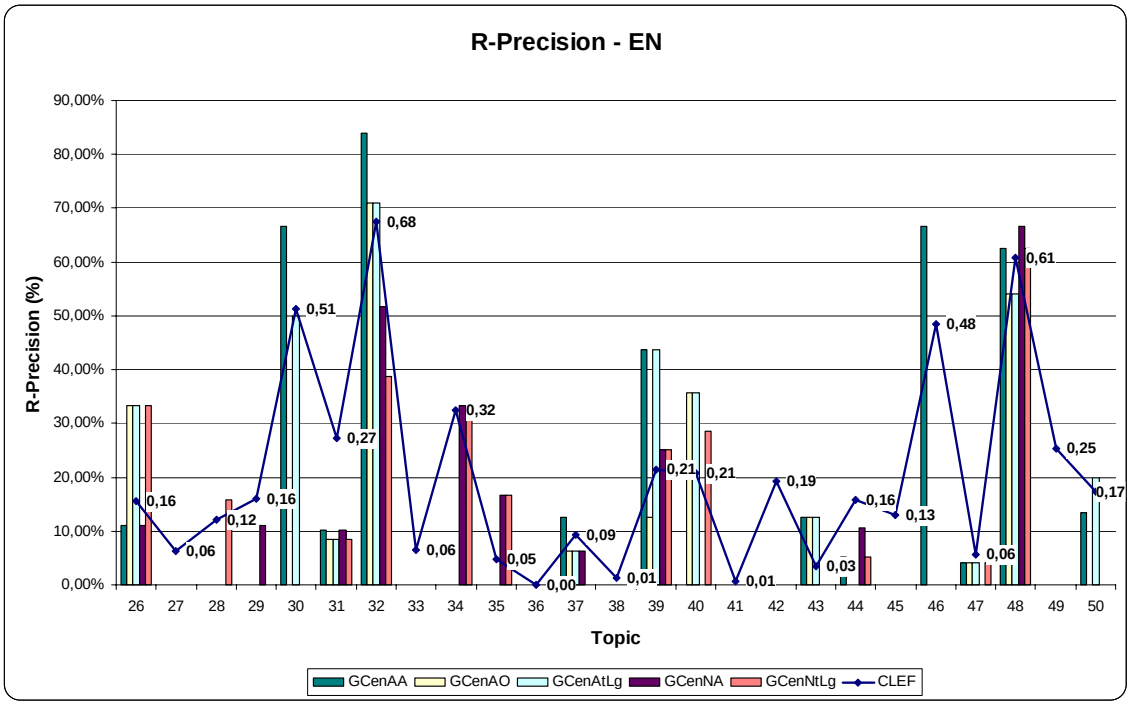
We tried a wide set of experiments, running several combinations of the variants described in the previous section.

For each of the monolingual tracks, we show a table with the run identifier, the precision at 0 and 1 points of recall, the mean average precision, the R-precision and the percentage deviation (in mean average precision and R-precision) from best one obtained. The last two rows show the two best results when the queries that satisfied textual and geographical sub-query are taking into account.

Results for Monolingual English

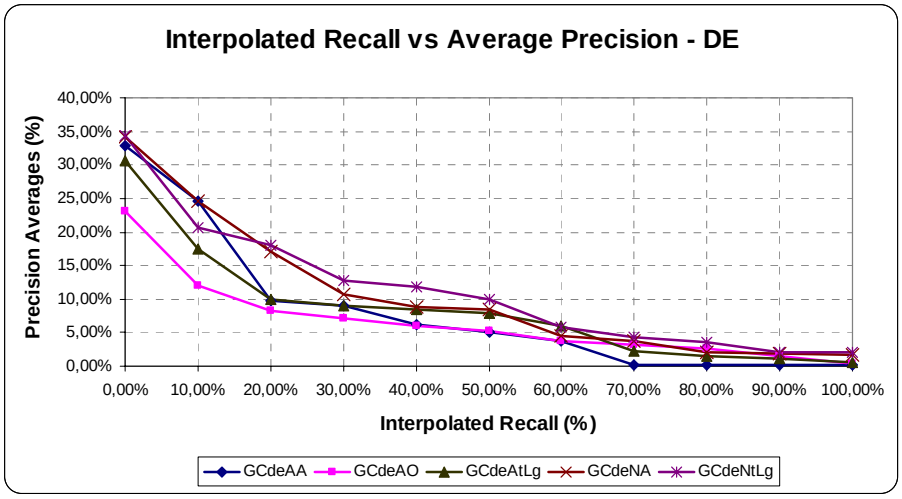
	AvP_0	AvP_1	MAP	R-Prec	MAP_diff	R-Prec_diff
GCenAA	0,29*	0,27*	0,14*	0,16*	6,15%	4,39%
GCenAtLg	0,25	0,24	0,13	0,14	6,70%	6,52%
GCenNA	0,28	0,25	0,09	0,10	10,82%	10,39%
GCenNtLg	0,24	0,22	0,09	0,11	10,38%	9,22%
GCenAO	0,19	0,17	0,09	0,10	10,84%	10,57%
GCenAA	0,42	0,40	0,20	0,23	-0,25%	-3,00%
GCenNA	0,47	0,42	0,15	0,16	4,86%	3,93%

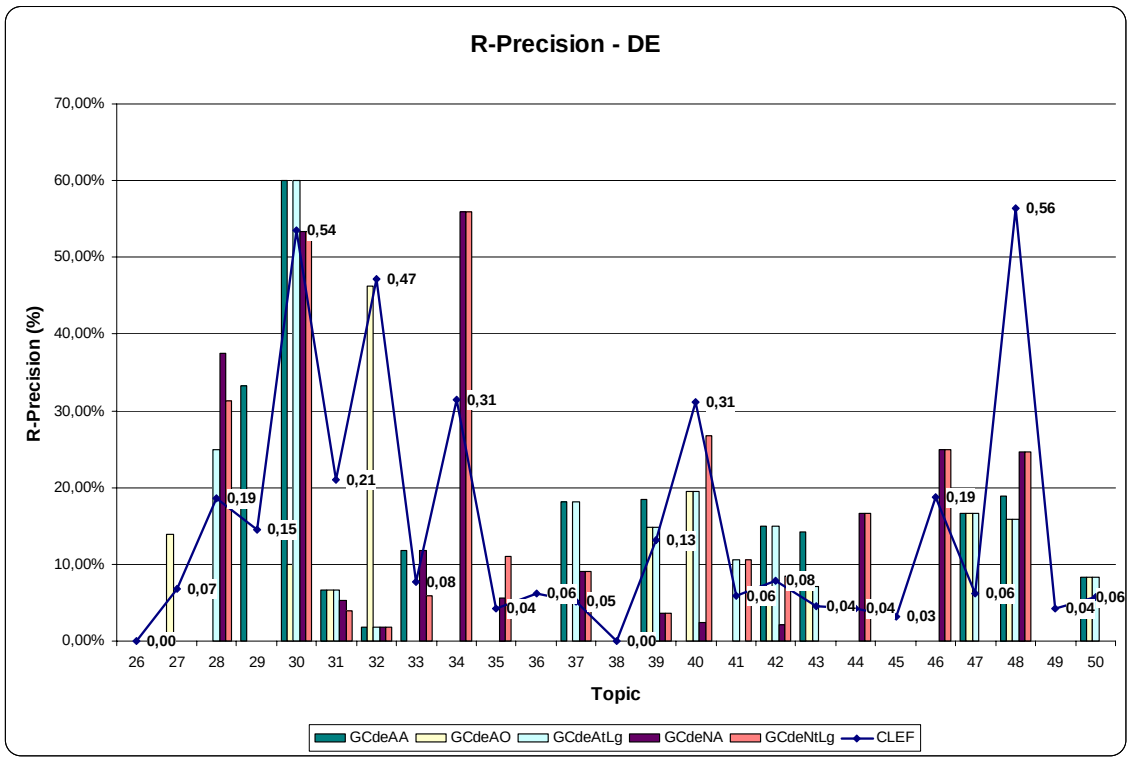




Results for Monolingual German

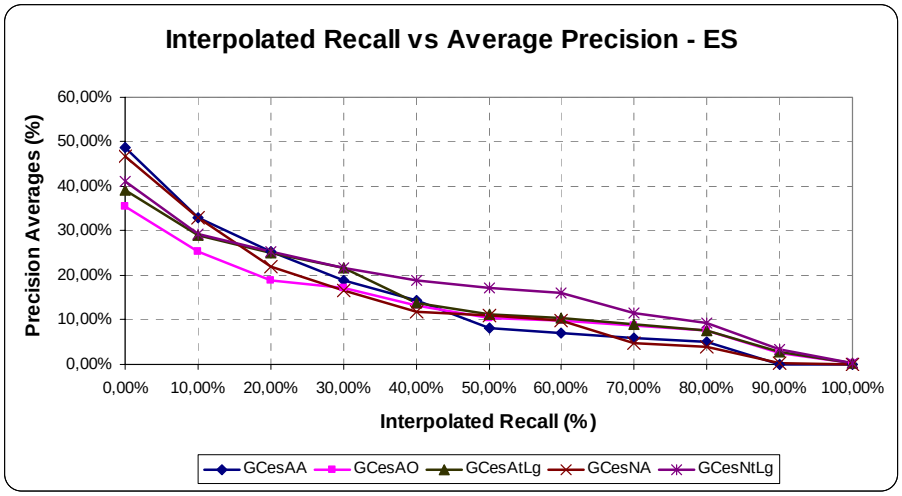
	AvP_0	AvP_1	MAP	R-Prec.	MAP_diff	R-Prec_diff
GCdeNtLg	0,34*	0,21	0,10	0,12*	4,17%	3,56%
GCdeNA	0,34*	0,25*	0,09	0,10	4,90%	4,90%
GCdeAA	0,33	0,25*	0,07	0,09	7,03%	6,16%
GCdeAtLg	0,31	0,18	0,07	0,09	6,83%	6,31%
GCdeAO	0,23	0,12	0,05	0,07	8,71%	8,47%
GCdeNA	0,50	0,36	0,14	0,15	0,53%	0,10%
GCdeNtLg	0,41	0,25	0,12	0,14	2,26%	1,37%

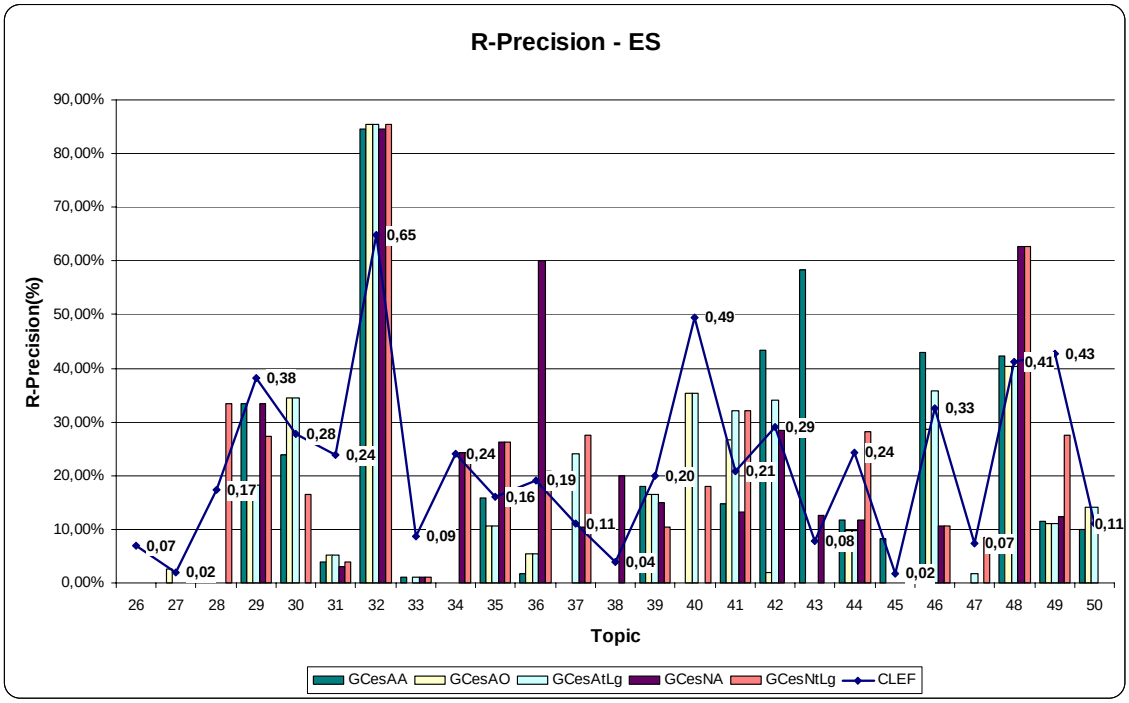




Results for Monolingual Spanish

	AvP_0	AvP_1	MAP	R-Prec	MAP_diff	R-Prec_diff
GCesNtLg	0,41	0,29	0,16*	0,19*	2,98%	3,50%
GCesAtLg	0,39	0,29	0,14	0,17	4,97%	5,51%
GCesAA	0,49*	0,33*	0,13	0,17	5,62%	5,08%
GCesNA	0,47	0,33	0,13	0,17	6,36%	4,90%
GCesAO	0,35	0,25	0,12	0,14	6,88%	8,26%
GCesNA	0,69	0,49	0,19	0,25	0,37%	-3,18%
GCesAA	0,64	0,43	0,18	0,22	1,37%	-0,30%





6. Conclusions and future works

If we compare the results achieved to the last campaign's results, we can assert that mean average precision gets worse when the textual geo-entity references are replaced with geographical tags. Part of this worsening is due to our experiments return cero pertinent documents if no documents satisfy the geographical sub-query. If we only analyze the results of queries that satisfied both textual and geographical terms, we observe that the designed experiments retrieval pertinent documents quickly, improving R-Precision values. Therefore we can conclude that the underlying idea of designed experiments apparently produces good results.

The advantage of basing the experiments on the combination of independent tools allows analyzing the efficiency of each phases of the recovery process and to work on which are weaker. The fact that the designed system presents great sensitivity to textual geo-references shows that one of the weak points of the system is the Named Entity Recognition, reason why our efforts will have to focus on exploring and evaluating different technique-based NER systems.

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; and by the Madrid's R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for Madrid Community), S-0505/TIC/000267.

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, Antonio Moreno-Sandoval and César de Pablo-Sánchez.

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 10/08/2006].

[3] CLEF 2005 Multilingual Information Retrieval resources page. On line http://www.computing.dcu.ie/~gjones/CLEF2005/Multi-8/ [Visited 10/08/2006].

[4] 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.

[5] Lana-Serrano, S.; Goñi-Menoyo, J.M.; and González-Cristóbal, J.C. MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[6] Lana-Serrano, S.; Goñi-Menoyo, J.M.; and González-Cristóbal, J.C. MIRACLE's 2005 Approach to Geographical Information Retrieval. Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 2005.

[7] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 10/08/2006].

[8] Robertson, S.E. et al. Okapi at TREC-3. In Overview of the Third Text REtrieval Conference (TREC-3). D.K. Harman (Ed.). Gaithersburg, MD: NIST, April 1995.

[9] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 10/08/2006].

[10] U.S. Geological Survey. On line http://www.usgs.gov [Visited 10/08/2006].

[11] U.S. National Geospatial Intelligence Agency. On line http://www.nga.mil [Visited 10/08/2006].