Report of MIRACLE team for the Ad-Hoc track in CLEF 2007

José Miguel Goñi-Menoyo[1], José C. González-Cristóbal[1, 3]
Julio Villena-Román[2, 3], Sara Lana-Serrano[1]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.

josemiguel.goni@upm.es, josecarlos.gonzalez@upm.es,
julio.villena@uc3m.es, sara.lana@upm.es

Abstract

This paper presents the 2007 MIRACLE's team approach to the AdHoc Information Retrieval *track*. The work carried out for this campaign has been reduced to monolingual experiments, in the standard and in the robust tracks. No new approaches have been attempted in this campaign, following the procedures established in our participation in previous campaigns.

For this campaign, runs were submitted for the following languages and tracks:
- Monolingual: Bulgarian, Hungarian, and Czech.
- Robust monolingual: French, English and Portuguese.

There is still some room for improvement around multilingual named entities recognition.

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]

Keywords

Linguistic Engineering, Information Retrieval, Trie Indexing, more keywords

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 fifth participation in CLEF. As well as monolingual and robust multilingual tasks, the team has participated in the ImageCLEF, Q&A, and GeoCLEF tracks.

The MIRACLE Information Retrieval toolbox is made of basic components: stemming, transformation (transliteration, elimination of diacritics and conversion to lowercase), filtering (elimination of stop and frequent words), proper nouns detection and extracting, and paragraph extracting, among others. Some of these basic components can be used in different combinations and order of application for document indexing and for query processing. Through our participation in previous campaigns, the integration procedure of the different modules is stable and, to some point, optimized.

MIRACLE makes use of its own indexing and retrieval engine, which is based on the *trie* data structure 0. Tries have been successfully used by the MIRACLE team for years, as an efficient storage and retrieval of huge lexical resources, combined with a continuation-based approach to morphological treatment [6]. For this campaign, runs were submitted for the following languages and tracks:

- Monolingual: Bulgarian, Hungarian, and Czech.
- Robust monolingual: French, English and Portuguese.

2 Description of the MIRACLE Toolbox

MIRACLE toolbox has already been described in previous campaigns papers [2], [3], [7]. We will say here that document collections and topics were pre-processed before feeding the indexing and retrieval engine, using different combinations of elementary processes. We will repeat here some relevant facts about these:

- **Extraction:** The extraction treatment has a special filter for extracting topic queries in the case of the use of the narrative field: some patterns that were obtained from the topics of the past campaigns are eliminated, since they are recurrent and misleading in the retrieval process. For example, for English, we can mention patterns as *"… are not relevant."*, or *"…are to be excluded"*. All the sentences that contain such patterns are filtered out.

- **Paragraphs extraction:** We have not used paragraph indexing this year, since the results we have obtained in this campaign and past ones have been disappointing.

- **Tokenization:** This process extracts basic text components, detecting and isolating punctuation symbols. Some basic entities are also treated, such as numbers, initials, abbreviations, years, and some proper nouns (see next item). The outcomes of this process are only single words, years that appear as numbers in the text (e.g. 1995, 2004, etc.), or entities.

- **Entities:** We consider that entities detection and normalization plays a central role in Information Retrieval, but it is a difficult task. For this year we have integrated a special module in the tokenization process that detects and marks some entities that have been previously collected from several sources into a lexical database for entities. These entities, which can be people names, place names, initials, abbreviations, etc., can consist of one or more words and special symbols, and their correct treatment is integrated into the tokenizer. For now, no entity normalization is done, so the same entity can appear in different forms and these are treated as different entities.

- **Filtering:** *Stopwords* lists in the target languages were initially obtained from [11], but were extended using several other sources and our own knowledge and resources. We have also compiled other lists of words to exclude from the indexing and querying processes, which were obtained from the topics of past CLEF editions and from our own background. We consider that such words have no semantics in the type of queries used in CLEF. As example, we can mention some of the English list: *find, appear, relevant, document, report,* etc.

- **Transformation:** The items that resulted from tokenization were normalized by converting all uppercase letters to lowercase, and accents eliminated. This has not been done for Bulgarian.

- **Stemming:** We used standard stemmers from Porter [8] for English, and from Neuchatel [11] for Hungarian, Bulgarian and Czech.

- **Indexing:** When all the documents processed through a combination of the former steps are ready for indexing, they are fed into our indexing *trie* engine to build the document collection index.

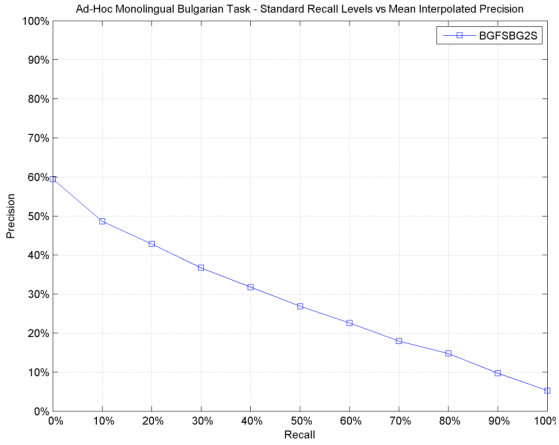
- **Retrieval:** When all the documents processed by a combination of the former steps are topic queries, they are fed to an ad-hoc front-end of the retrieval *tri*e engine to search the previously built document collection index. In the 2007 experiments, only OR combinations of the search terms were used. The retrieval model used is the well-known Robertson's Okapi BM-25 [9] formula for the probabilistic retrieval model, without relevance feedback.

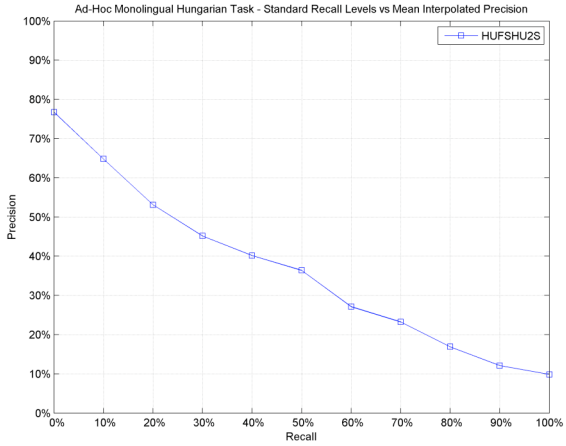
3 Results for the monolingual and robust tasks

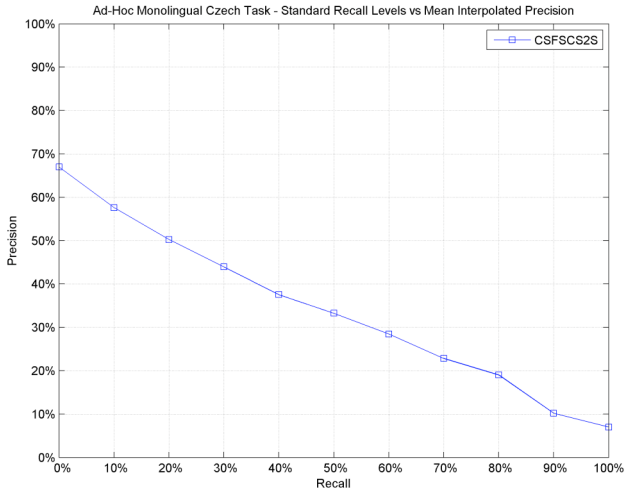
The following table and graphic representation summarize the performance of our official experiments in the monolingual tasks (using the topic fields title/description).

Precision figures for monolingual experiments

lang	Average Precision	Prec. at 0	Prec. at 100
bg	0.2717	0.5946	0.0531
cz	0.3203	0.6697	0.0701
hu	0.3499	0.7672	0.987



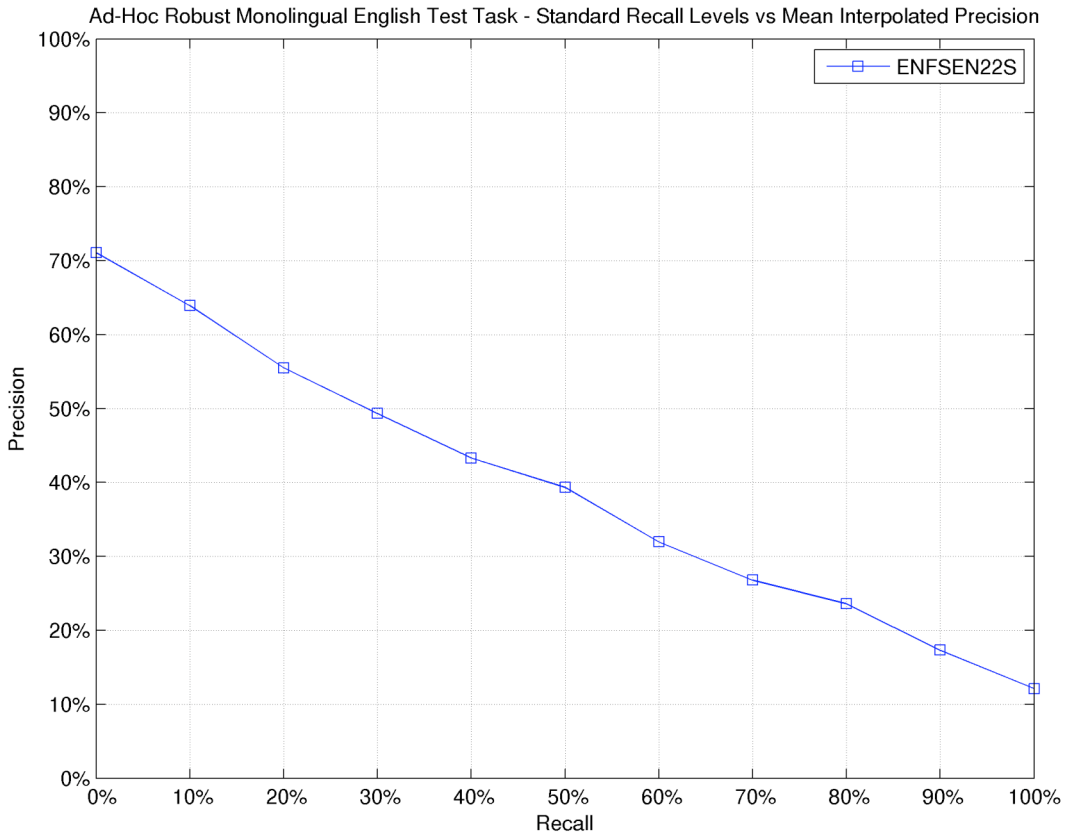




In the case of the monolingual robust task, only the results for English will be shown, as our PT and FR runs did not match the interpretation made in the assessment concerning the available collections and topics included in each experiment. In the results, the mean average precision figures are given.

Precision figures for robust monolingual experiments

lang	Average Precision	Prec. at 0	Prec. at 100
en	0.3778	0.7109	0.1211



4 Conclusions and future work

This year we have not changed our previous processing scheme, using the same improvements incorporated last year regarding proper nouns and entities detection and indexing. For this reason, the obtained results must be quite similar to previous ones. The only element that makes the processing of each language different has to do with the stemming components and stopword lists.

It is clear that the quality of the tokenization step is of paramount importance for precise document processing. We still think that a high-quality entity recognition (proper nouns or acronyms for people, companies, countries, locations, and so on) could improve the precision and recall figures of the overall retrieval, as well as a correct recognition and normalization of dates, times, numbers, etc. Although we have introduced some improvements in our processing scheme during the last years, a good multilingual entity recognition and normalization tool is still missing.

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 other 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] Goñi-Menoyo, J.M.; González-Cristóbal, J.C.; and Villena-Román, J. MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. 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).

[3] Goñi-Menoyo, J.M.; González, J.C.; and Villena-Román, J. Miracle's 2005 Approach to Monolingual Information Retrieval. Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 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. Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers (Carol Peters, Paul Clough, Julio Gonzalo, et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005.

[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] González, J.C.; Goñi-Menoyo, J.M.; and Villena-Román, J. Miracle's 2005 Approach to Cross-lingual Information Retrieval. Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 2005.

[8] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 18/07/2006].

[9] 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.

[10] Savoy, Jacques. Report on CLEF-2003 Multilingual Tracks. 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. 64-73. Springer, 2004.

[11] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 18/07/2006].