Miracle's 2005 Approach to Monolingual Information Retrieval

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

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

josemiguel.goni@upm.es, jgonzalez@dit.upm.es,
julio.villena@uc3m.es

Abstract

This paper presents the 2005 Miracle's team approach to Monolingual Information Retrieval. The goal for the experiments in this year was twofold: continue testing the effect of combination approaches on information retrieval tasks, and improving our basic processing and indexing tools, adapting them to new languages with *strange* encoding schemes. The starting point was a set of basic components: stemming, transforming, filtering, proper nouns extracting, paragraph extracting, and pseudo-relevance feedback. Some of these basic components were used in different combinations and order of application for document indexing and for query processing. Second order combinations were also tested, by averaging or selective combination of the documents retrieved by different approaches for a particular query.

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, BM25 Probabilistic Retrieval, Merging Experiments

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 the third participation in CLEF, after years 2003 and 2004 [2], [3], [5], [6], [7], [8], [9], [13], [14]. As well as bilingual, monolingual and cross lingual tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and GeoCLEF tracks.

The starting point was a set of basic components: stemming, transformation (transliteration, elimination of diacritics and conversion to lowercase), filtering (elimination of stop and frequent words), proper nouns extracting, paragraphs extracting, and pseudo-relevance feedback. Some of these basic components are used in different combinations and order of application for document indexing and for query processing. Second order combinations were also tested, mainly by averaging or by selective combination of the documents retrieved by different approaches for a particular query. When evidence is found of better precision of one system at one extreme of the recall level (i.e. 1), complemented by the better precision of another system at the other recall end (i.e. 0), then both are combined to benefit from their complementary results.

In addition, during the last year our group has been improving an indexing system based on the trie data structure, which was reported last year [2]. Tries [1] 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 [4]. However, the adaptation of these structures to manage efficiently document indexing and retrieval for IR applications has been a hard task, mainly in the issues concerning the performance of the construction of the index. Thus, this year we have used only our trie-based indexing system, and so, the Xapian[1] [15] indexing system used in the previous CLEF editions was no more needed.[2] In fact, we have been able to make more experiments than the previous year, since we have had more time available.

For the 2005 monolingual campaign, runs were submitted for the following languages: Bulgarian, French, Hungarian, and Portuguese.

2 Description of the MIRACLE Toolbox

Document collections were pre-processed before indexing, using different combinations of elementary processes, each one oriented to a particular experiment. For each of these, topic queries were also processed using the same combination of processes. (Although some variants have been used, as will be described later.)

The baseline approach to document and topic query processing is made up of a combination of the following steps:

- **Extraction:** The raw text from different document collections or topic files is extracted with ad-hoc scripts that selected the contents of the desired XML elements. All those permitted for automatic runs were used. (Depending on the collection, all of the existing TEXT, TITLE, LEAD1, TX, LD, TI, or ST for document collections, and the contents of the TITLE, DESC, and NARR for topic queries.) The contents of these tags were concatenated, without further distinction to feed subsequent processing steps. This 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:** In some experiments, we indexed paragraphs[3] instead of documents. Thus, the subsequent retrieval process returned document paragraphs, so we needed to combine the relevance measures from all paragraphs retrieved for the same document. We tested several approaches for this combination, for example counting the number of paragraphs, adding relevance measures or using the maximum of the relevance figures of the paragraphs retrieved. Experimentally, we got best results using the following formula for document relevance:

$$rel_N = rel_{mN} + \xi \cdot \frac{1}{n} \cdot \sum_{j \neq m} rel_{jN} ,$$

  where n is the number of paragraphs retrieved for document N, rel_{iN} is the relevance measure obtained for the i-th paragraph of document N, and m refers to the paragraph with maximum relevance. The coefficient ξ was adjusted experimentally to 0.75. The idea behind this formula is to give paramount importance to the maximum paragraph relevance, but taking into account the rest of the relevant paragraphs to a lesser extent. Paragraph extraction has not been used for topics processing.
- **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.).
- **Filtering:** All words recognized as *stopwords* are filtered out. *Stopwords* in the target languages were initially obtained from [12], but were extended using several other sources and our own knowledge and

[1] Xapian has rendered an excellent service to Miracle group in the 2003 and 2004 editions of CLEF. It is robust, quite efficient, and was well suited to our purposes.

[2] The main reason for changing the indexing engine is efficiency: as an example, we reduced the time needed to indexing one of the experiments scheduled from days to hours. In addition to that, our trie-based indexing system was able to index collection of texts encoded in UTF-8, so we have had no need to use transliteration schemes for languages such as Bulgarian.

[3] Paragraphs are either marked by the <P> tag in the original XML document, or are separated from each other by two carriage returns, so they are easily detected.

resources. We have also other lists of words to exclude from the indexing and querying processes, which were obtained from the topics of past CLEF editions. 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: *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 process is usually done after stemming, although it can be done before, though the resulting lexemes are different. We ought to make it before stemming in the case of the Bulgarian and Hungarian languages, since these stemmers did not work well with uppercase letters. Note that the accent removal process is not applicable for Bulgarian language.
- **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. We used standard stemmers from Porter [10] for most languages, except for Hungarian, where we used a stemmer from Neuchatel [12].
- **Proper noun extraction:** In some experiments, we try to detect and extract proper nouns in the text. The detection was very simple: Any chunk that outcomes from a the tokenization process is considered a proper noun provided that its firs letter is uppercase, unless such word is included in the *stopwords* list or in a specifically built list of words that are not suitable to be proper nouns (mainly verbs and adverbs). We opted for this simple strategy[4] since we had not available huge lists of proper nouns. In the experiments that used this process, only the proper nouns extracted from the topics fed a query to an index of documents of *normal* words, where neither proper nouns were extracted nor stemming was done.
- **Final use:**
    - **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 2005 experiments, only OR combinations of the search terms were used. The retrieval model used is the well-known Robertson's Okapi BM-25 [11] formula for the probabilistic retrieval model, without relevance feedback.

After retrieval, some other special processes were used to define additional experiments:

- **Pseudo-relevance feedback:** We used this technique in some experiments. After a first retrieval step, we processed the first retrieved document to get their indexing terms that, after a standard processing[5] (see below) are fed back to a second retrieval step, whose result is used.
- **Combination:** The results from some basic experiments can be combined in different ways. The underlying hypothesis is that, to some extent, the documents with a good score in almost all experiments are more likely to be relevant than other documents that have a good score in one experiment but a bad one in others. Two strategies were followed for combining experiments:
    - **Average:** The relevance figures obtained using the probabilistic retrieval in all the experiments to be combined for a particular document in a given query is added. This approach combines the relevance figures of the experiments without highlighting a particular experiment.
    - **Asymmetric WDX combination:** In this particular type of combination, two experiments are combined in the following way: The relevance of the first D documents for each query of the first experiment is preserved for the resulting combined relevance, whereas the relevance for the remaining documents in both experiments are combined using weights W and X. We have only run experiments labeled "011", that is, the ones that get the document more relevant from the first basic experiment and all the remaining documents retrieved from the second basic experiment, re-sorting all the results using the original relevance measure value.

[4] Note that multi-word proper nouns cannot be treated this way.

[5] Both retrieval processes can be independent from each other: we could have used two different treatments for the queries and documents, so using different indexes for each of the retrievals. In our case, only standard treatments were used for both retrieval steps.

3 Basic experiments

For this campaign we have designed several basic experiments in which the documents for indexing and the topic queries for retrieval are processed using the same combination of the steps described in the previous section. For ease of reference we used letters that denote a set of these basic processes or variations of the extraction process for topic queries:

- **S:** Standard or baseline treatment: tokenization, filtering, stemming, and transformation[6].
- **N:** Non-stemming treatment: tokenization, filtering, and transformation.
- **R:** Use the narrative field in the topics.
- **T:** Do not use narrative field.

So, the basic experiments are denoted **SR, ST, NR,** or **NT**. We designed additional experiments:

- **P:** Treatment that extracts proper nouns from topic queries, even using the narrative field. Since no stemming is made on the proper nouns detected, a non-stemmed index is needed. Thus, the only possible experiment is denoted **NP**.
- **H:** The standard treatment (**S**) is made on the index built from the paragraphs of the documents from the collection. The combination of the paragraphs retrieved is done as stated in the previous section. Depending on the fields selected in the topic queries, we can denote these experiments as **HR** or **HT**.
- **r1:** Pseudo-relevance feedback, as described above. As we only used **SR** runs to feed this type of run, we denoted it as **r1SR**.

Finally, we describe the combining experiments we used:

- **x<run1>WD<run2>X:** This denotes an asymmetric DWX combination from experiments <run1> and <run2>. For example, we run experiment xNP01HR1, which gets the first document retrieved from the NP run, with its original relevance value and all the documents retrieved in the experiment HR also with their original relevance values, then resorting all these relevance figures.
- **a<run1>….<runZ>:** This denotes the average run for all the Z runs stated. For example, we run the experiment denoted aHTSTxNP01ST1, which adds all the relevance measure values obtained for each document in the experiments HT, ST, and xNP01ST1.

4 MIRACLE results for CLEF 2005

We tried a wide set of experiments, running several combinations of the variants described in the previous section, although not all the possible tried due to evident limitations of computing resources and time. The experiments were tried trying to test a wider and richer set of trials.

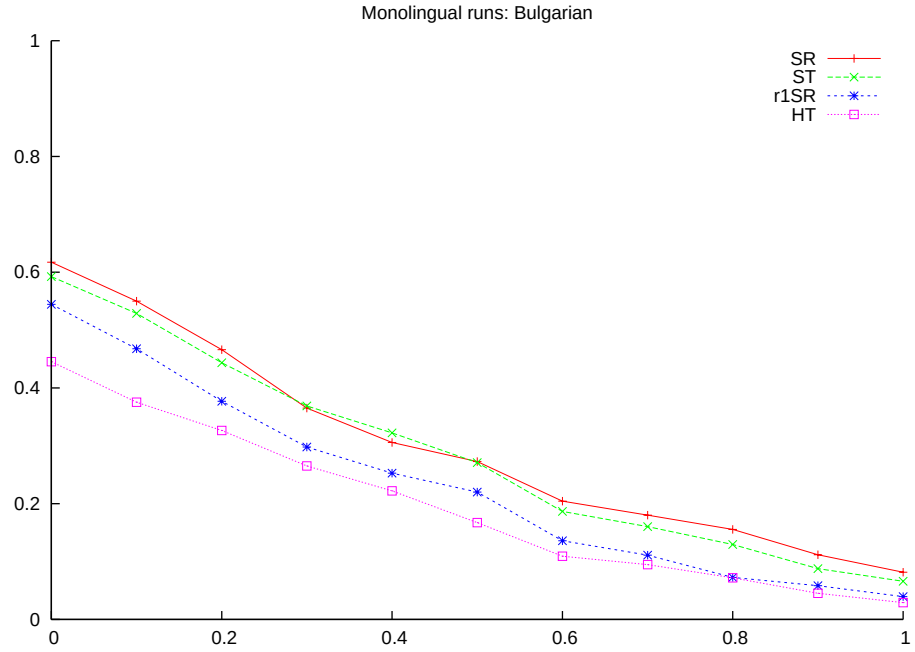
To compare these approaches, we used these techniques following the instructions given for CLEF 2004 (corpora and topic queries) and using the appropriate *qrels* available at the beginning of this campaign. The experiments that provided the best precision results in the CLEF 2004 scenario were selected for submission to CLEF 2005.

The results from our experiments follow. For each of the monolingual tracks, we show a table with the precision at 0 and 1 points of recall, the average precision, the percentage deviation (in average precision) from best one obtained, and the run identifier. The results are sorted in average precision ascending order, but an asterisk marks all the best precision values for each column. The submitted runs to CLEF 2005 are shown in boldface, and the figures show the precision-recall graphs for the submitted runs as well as our best[7] runs, provided that these were not submitted.

[6] As it was commented before, in some cases these two steps are made in reverse order.
[7] Be the best run in average precision, or in precision at 0 or 1 points of recall.

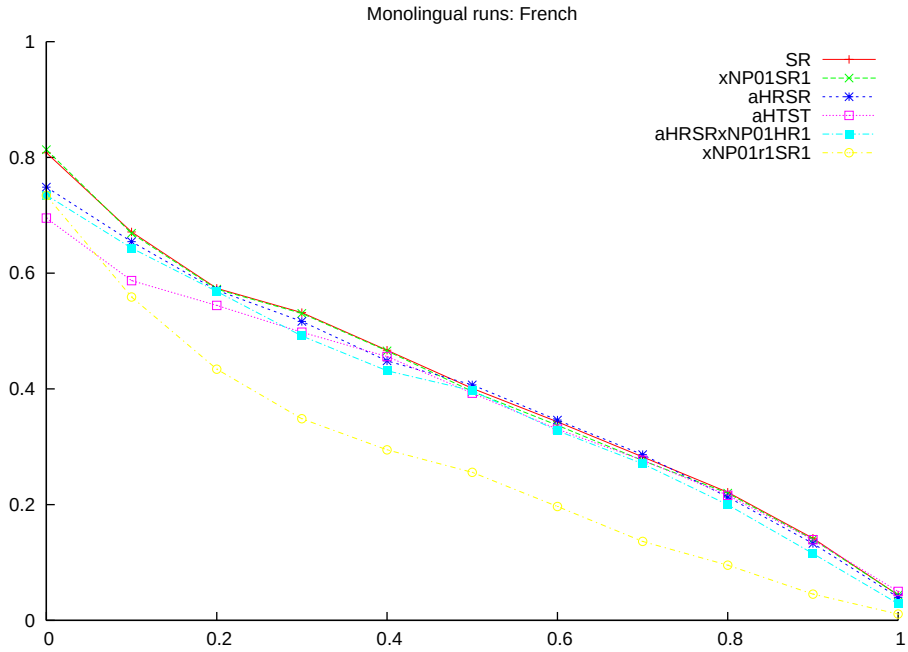
Results for monolingual Bulgarian

At 0	At 1	Avgp	%	Run
0.4453	**0.0291**	**0.1804**	**-36.01%**	**HT**
0.4419	0.0315	0.1886	-33.10%	HR
0.5445	**0.0393**	**0.2191**	**-22.28%**	**r1SR**
0.5174	0.0678	0.2200	-21.96%	NR
0.5552	0.0665	0.2328	-17.42%	NT
0.5924	**0.0659**	**0.2676**	**-5.07%**	**ST**
0.6173*	0.0816*	0.2819*	-0.00%	SR



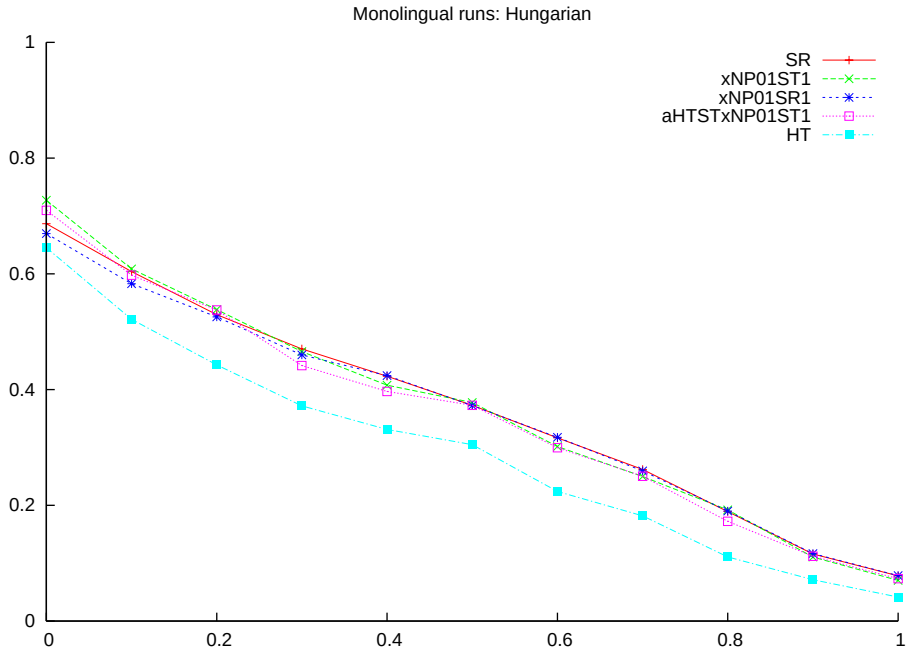
Results for monolingual French

At 0	At 1	Avgp	%	Run
0.3473	0.0180	0.1254	-68.02%	NP
0.7341	**0.0109**	**0.2636**	**-32.77%**	**xNP01r1SR1**
0.7338	0.0190	0.2713	-30.81%	r1SR
0.6916	0.0227	0.3157	-19.48%	NT
0.7071	0.0252	0.3298	-15.89%	xNP01HR1
0.7046	0.0251	0.3350	-14.56%	HR
0.7952	0.0279	0.3431	-12.50%	NR
0.7347	**0.0289**	**0.3675**	**-6.27%**	**aHRSRxNP01HR1**
0.6951	0.0501*	0.3692	-5.84%	ST
0.6951	0.0501*	0.3692	-5.84%	HT
0.6951	**0.0501***	**0.3692**	**-5.84%**	**aHTST**
0.7486	**0.0398**	**0.3833**	**-2.24%**	**aHRSR**
0.8133*	0.0437	0.3883	-0.97%	xNP01SR1
0.8081	0.0437	0.3921*	-0.00%	SR



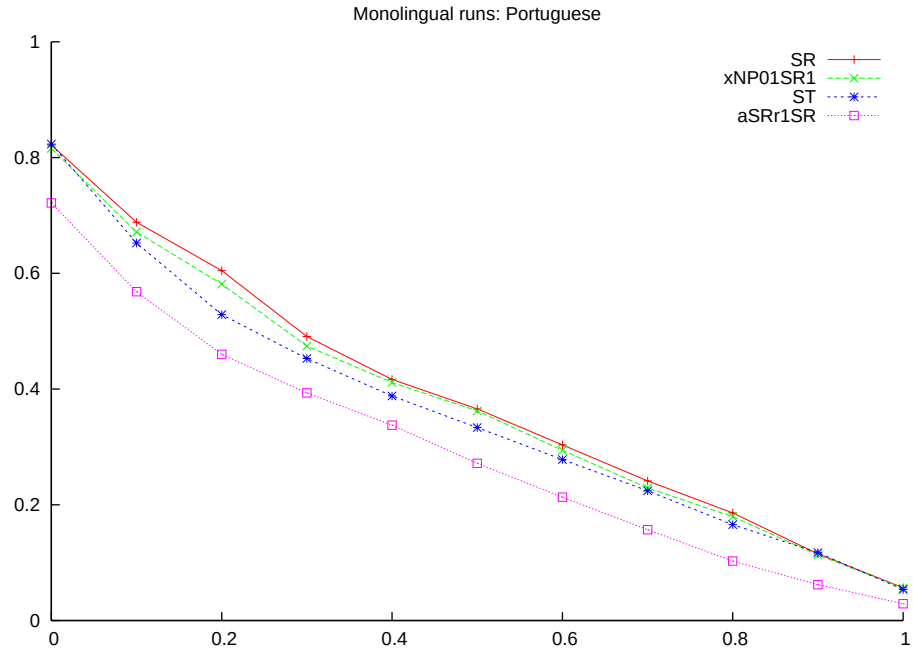
Results for monolingual Hungarian

At 0	At 1	Avgp	%	Run
0.4484	0.0534	0.1776	-49.77%	NP
0.5397	0.0578	0.2263	-36.00%	NT
0.6079	0.0531	0.2641	-25.31%	NR
0.6225	0.0421	0.2721	-23.05%	xNP01HT1
0.6452	**0.0414**	**0.2770**	**-21.66%**	**HT**
0.6007	0.0426	0.2777	-21.46%	xNP01HR1
0.6447	0.0413	0.2843	-19.60%	HR
0.6668	0.0562	0.3085	-12.75%	aHTSTxNP01HT1
0.6660	0.0651	0.3266	-7.64%	aHRSRxNP01HR1
0.7054	0.0677	0.3373	-4.61%	aHRSR
0.7096	**0.0733**	**0.3435**	**-2.86%**	**aHTSTxNP01ST1**
0.7197	0.0703	0.3493	-1.22%	ST
0.6697	0.0785*	0.3501	-0.99%	xNP01SR1
0.7272*	**0.0703**	**0.3520**	**-0.45%**	**xNP01ST1**
0.6867	0.0781	0.3536*	-0.00%	SR



Results for monolingual Portuguese

At 0	At 1	Avgp	%	Run
0.4586	0.0396	0.1669	-54.87%	NP
0.6718	0.0070	0.2214	-40.13%	xNP01r1SR1
0.7035	0.0232	0.2358	-36.24%	r1SR
0.7217	**0.0290**	**0.2832**	**-23.42%**	**aSRr1SR**
0.7723	0.0455	0.2957	-20.04%	NT
0.8074	0.0469	0.3198	-13.52%	NR
0.8232*	**0.0536**	**0.3456**	**-6.54%**	**ST**
0.8160	**0.0561**	**0.3628**	**-1.89%**	**xNP01SR1**
0.8217	**0.0566***	**0.3698***	**-0.00%**	**SR**



5 Conclusions

Except for Portuguese, the best results obtained came from runs that were not submitted, since we obtained worse results using the 2004 queries and *qrels* than with the submitted ones. We think that this behavior can be explained since the results depend to a great extent on the different topics selected each year. It is worth to note that we obtained the best results using the narrative field of the topic queries in all cases, as well as the standard processing approach (SR runs).

We expected to have had better results using combinations of proper nouns indexing with standard (SR or ST) runs, as it seemed to follow from the results from 2004 campaign, but it has not been the case. 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. Pseudo-relevance feedback has not performed quite well, but we run quite few experiments to extract general conclusions. On the other hand, these runs had a lot of querying terms, what made them very slow.

Regarding the basic experiments, the general conclusions were known in advance: retrieval performance can be improved by using stemming, filtering of frequent words and appropriate weighting.

6 Future work

Future work of the MIRACLE team in these tasks will be directed to several lines of research: (a) Tuning our indexing and retrieval *trie*-based engine in order to get even better performance in the indexing and retrieval phases, and (b) improving the tokenization step; in our opinion, this is one of the most critical processing ones and can improve the overall results of the IR process. A good entity recognition and normalization is still missing in our processing scheme for these tasks. We need better performance of the retrieval system to drive runs that are efficient when the query has some hundred terms, as occurs when using pseudo-relevance feedback. We need also to explore further the combination schemes with these enhancements to the basic processes.

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, Sara Lana-Serrano, José Luis Martínez-Fernández, Paloma Martínez-Fernández, Ángel Martínez-González, 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, 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).

[3] 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.

[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] 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).

[6] 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.

[7] 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.

[8] 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.

[9] 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).

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

[11] 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.

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

[13] 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.

[14] 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.

[15] Xapian: an Open Source Probabilistic Information Retrieval library. On line http://www.xapian.org [Visited 13/07/2005].