MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining Without Using a Single Approach

José M. Goñi-Menoyo[1], José C. González-Cristóbal[1,3], and 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 the Ad-Hoc Information Retrieval tasks. The goal for the experiments this year was twofold: to 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 extraction, paragraph extraction, 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. In the multilingual track, we concentrated our work on the merging process of the results of monolingual runs to get the overall multilingual result, relying on available translations. In both cross-lingual tracks, we have used available translation resources, and in some cases we have used a combination approach.

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 third participation in CLEF, after 2003 and 2004. As well as bilingual, monolingual and cross lingual tasks, the team has participated this year 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), extracting proper nouns, extracting paragraphs, 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

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 44 – 53, 2006.

other recall end (i.e. 0), then both are combined to benefit from their complementary results.

Additionally, during the last year our group has been improving an indexing system based on the trie data structure, which was reported last year. Tries [1] have been successfully used by the MIRACLE team for years, as an efficient technique for the 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 document indexing and retrieval for IR applications efficiently 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 [12] indexing system used in the previous CLEF editions was no longer needed. In fact, we have been able to carry out more experiments than the previous year, since we have had more computing time available because of this improvement in indexing efficiency.

For the 2005 bilingual track, runs were submitted for the following language pairs: English to Bulgarian, French, Hungarian and Portuguese; and Spanish to French and Portuguese. For the multilingual track, runs were submitted using as source language English, French, and Spanish. Finally, in the monolingual case runs were submitted for 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 towards 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, *"… are not relevant."*, or *"…are to be excluded."*. All the sentences that contain these patterns are filtered out.
- **Paragraphs extraction:** In some experiments, we indexed paragraphs[1] instead of documents. Thus, the subsequent retrieval process returned document paragraphs, so we needed to combine the relevance measures from all paragraphs retrieved for

[1] 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.

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_{jN} is the relevance measure obtained for the j-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 was not used for topic 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 [11], but were extended using several other sources and our own knowledge and resources. We also used 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; for example, in 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 eliminating accents. This process is usually carried out after stemming, although it can be done before, but the resulting lexemes are different. We ought to do 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.
- **Stemming:** This process is applied to each one of the words to be indexed or used for retrieval. We used standard stemmers from Porter [8] for most languages, except for Hungarian, where we used a stemmer from Neuchatel [11].
- **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 results from the tokenization process is considered a proper noun provided that its first letter is uppercase, unless this 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[2] since we did not have 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 carried out.
- **Linguistic processing:** In the Multi-8 track, and only in the case of Spanish as topic language, we tested an approach consisting in pre-processing the topics with

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

a high quality morphologic analysis tool. This tool is STILUS[3]. STILUS not only recognizes closed words, but also expressions (prepositional, adverbial, etc.). In this case, STILUS is simply used to discard closed words and expressions from the topics and to obtain the main form of their component words (in most cases, singular masculine or feminine for nouns and adjectives and infinitive for verbs). The queries are so transformed to a simple list of words that are passed to the automatic translators (one word per line).

– **Translation:** For cross-lingual tracks, popular on-line translation or available dictionary resources were used to translate topic queries to target languages: ATRANS was used for the pairs EsFr and EsPt; Bultra and Webtrance for EnBg[4]; MoBiCAT for EnHu; and SYSTRAN was used for the language pairs EnFr, EsFr, and EnPt. However, for multilingual runs having English as topic language, we avoided working on the translation problem for some runs. In this case, we have used the provided translations for topic queries [2], testing Savoy's [10] approach to translation concatenations. Two cases were considered: all available translations are concatenated, and selected translations are concatenated. Table 1 shows the translations used for both cases.

  In the Multi-8 track we also used automatic translation systems: for Spanish and French as topic languages, ATRANS was used for the pairs EsFr and EsPt; WorldLingo for EsDe, EsIt, and EsNl; InterTrans for EsFi, EsSv, FrFi, and FrSv; and SYSTRAN was used for all the other language pairs. Only one translator was used for each pair.

– **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 aforementioned steps are topic queries, they are fed to an ad-hoc front-end of the retrieval *trie* 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 [9] 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.

---

[3] STILUS® is a trademark of DAEDALUS-Data, Decisions and Language, S.A. It is the core of the Spanish-processing tools of the company, that include spell, grammar and style checkers, fuzzy search engines, semantic processing, etc.

[4] In the case of Bulgarian, an average combination of the results from the translations with the Webtrance and Bultra systems from English to Bulgarian has also been used.

[5] Both retrieval processes can be independent of 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.

Table 1. Available automatic translations used for concatenating

	Topic language							
Translation	DE	EN	ES	FI	FR	IT	NL	SV
ALT					A			
BA1	AH		AH	AH	AH	AH	AH	AH
BA2	A		A	A	A	A	A	A
BA3	A		A	A		A	A	A
FRE	AH		AH		AH	AH	AH	
GOO	AH		AH		AH	AH		
INT	A		A	AH	A	A	A	AH
LIN					A			
REV	AH		AH		AH			
SYS	AH		A		A	A		

ALT for Babelfish Altavista, BA1, BA2, and BA3[6] for Babylon, FRE for FreeTranslation, GOO for Google Language Tools, INT for InterTrans, LIN for WordLingo, REV for Reverso, and SYS for Systran. The entries in the table contain A (for ALL) if a translation is available for English to the topic language shown in the heading row of a column, and it is used for the concatenation of all available translations; and H if a translation is used for the selected concatenation of translations.

– **Combination:** The results from some basic experiments were 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 are 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 most relevant documents 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.

– **Merging:** In the multilingual case, the approach used requires that the monolingual results list for each one of the target languages have to be merged. The results obtained are very sensitive to the merging approach for the relevance measures. The

[6] The digit after BA shows how many words are used from the translation of a word, provided that it returns more than one.

probabilistic BM25 [9] formula used for monolingual retrieval gives relevance measures that depend heavily on parameters that are too dependent on the monolingual collection, so it is not very good for this type of multilingual merging, since relevance measures are not comparable between collections. In spite of this, we carried out merging experiments using the relevance figures obtained from each monolingual retrieval process, considering three cases:[7]

- Using original relevance measures for each document as obtained from the monolingual retrieval process. The results are made up of the documents with greater relevance measures.
- Normalizing relevance measures with respect to the maximum relevance measure obtained for each topic query *i* (*standard normalization)*:

$$rel_{i\,norm} = \frac{rel_i}{rel_{i\,\max}}$$

  Then, the results are made up of the documents with greater normalized relevance measures.
- Normalizing relevance measures with respect to the maximum and minimum relevance measure obtained for each topic query *i (alternate normalization)*:

$$rel_{i\,alt} = \frac{rel_i - rel_{i\,\min}}{rel_{i\,\max} - rel_{i\,\min}}$$

  Then, the results are made up of the documents with greater alternate normalized relevance measures.

  In addition to all this, we tried a different approach to merging: Considering that the more relevant documents for each of the topics are usually the first ones in the results list, we will select from each monolingual results file a variable number of documents, proportional to the average relevance number of the first N documents. Thus, if we need 1,000 documents for a given topic query, we will get more documents from languages where the average relevance of the first N relevant documents is greater. We did all this both from non-normalized runs, but normalized after the merging process is carried out (with *standard* and *alternate* normalization); and from runs normalized with *alternate* normalization. We tested several cases using results from baseline runs, using several values for N: 1, 10, 50, 125, 250, and 1,000.

3 Description of the Experiments

For this campaign we have designed several experiments in which the documents for indexing and the topic queries for retrieval are processed using a particular combination of some of the steps described in the previous section. A detailed inventory of the experiments, the processes used for each one, and their encoding in the name of the experiment can be found in the papers submitted to the CLEF 2005 Workshop ([3], [5]). Details of the documents collections and the tasks can be found in the introduction [8] and track overview [6] papers.

[7] Round-robin merging for results of each monolingual collection has not been used.

Several hundreds of experiments were run, and the criterion for choosing the ones to be submitted was the runs that obtained best results using topic queries and *qrels* sets from the 2004 campaign. Except for Portuguese, the best results obtained came from runs that were not submitted. 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 noting that we obtained the best results using the narrative field of the topic queries in all cases, as well as the standard processing approach.

We expected to have had better results using combinations of proper noun indexing with standard 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 ran quite few experiments of this type to extract general conclusions. On the other hand, these runs had a lot of querying terms, which 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.

Regarding cross-lingual experiments, the MIRACLE team has worked on their merging and combining aspects, departing from the translation ones. Combining approaches seems to improve results in some cases. For example, the average combining approach allows us to obtain better results when combining the results from translations for Bulgarian than the Bultra or Webtrance systems alone. In multilingual experiments, combining (concatenating) translations permits better results, as was reported previously [10], when good translations are available. Regarding the merging aspects, our approach did not obtain better results than standard merging, whether normalized or not. Alternate normalizations seem to behave better than the standard normalization, whereas the latter behaves better than no normalization. This occurs too when normalization is used in our own approach to merging.

Regarding the approach consisting of preprocessing queries in the source topic language with high quality tools for extracting content words before translation, the results have been good when used in the case of Spanish (with our tool STILUS). This approach achieved the best precision figures at 0 and at 1 recall extremes, although worse average precision than other runs.

In the appendix we have included two figures that summarize these results. Figure 1 shows a comparison of the results obtained in the best runs in the monolingual experiments for each target language. The best results are obtained for French and Portuguese, and the worst for Bulgarian. Figure 2 shows the results obtained in the best runs in the cross-lingual experiments for bilingual and multilingual runs, considering all source languages used.

4 Conclusions and 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. Good entity recognition and normalization is still missing from 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 also need to explore further the combination schemes with these enhancements of the basic processes.

Regarding cross-lingual tasks, future work will be centered on the merging aspects of the monolingual results. The translation aspects of this process are of no interest to us, since our research interests depart from all this: we will only use translation resources available, and we will try to combine them to get better results.

On the other hand, the process of merging the monolingual results is very sensitive in the way it is done; there are some techniques to be explored. In addition to that, perhaps a different way of measuring relevance is needed for monolingual retrieval when multilingual merging has to be carried out. Such a measure should be independent of the collection, so monolingual relevance measures would be comparable.

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, J.-I., Morimoto, K., and Sato, T.: An Efficient Implementation of Trie Structures. Software Practice and Experience 22(9): 695-721 (1992)
2. CLEF 2005 Multilingual Information Retrieval resources page. On line http://www.computing.dcu.ie/~gjones/CLEF2005/Multi-8/ [Visited 11/08/2005].
3. 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) Online http://clef.isti.cnr.it/2005/working_notes/workingnotes2005/gonzalez05.pdf [Visited 05/11/2005].
4. Goñi-Menoyo, J. M., González-Cristóbal, J. C., and Fombella-Mourelle, J.: An optimised trie index for natural language processing lexicons. MIRACLE Technical Report. Universidad Politécnica de Madrid (2004)

5. 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) On line http://clef.isti.cnr.it/2005/working_notes/workingnotes2005/menoyo05.pdf [Visited 05/11/2005].
6. Di Nunzio, G. M., Ferro, N., and Jones, G. J. F.: CLEF 2005: Ad Hoc Multilingual Track Overview. Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 (in this volume).
7. Peters, C.: What happened in CLEF 2005. Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 (in this volume).
8. Porter, M.: Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 13/07/2005].
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 (1995)
10. Savoy, J.: 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 13/07/2005].
12. Xapian: an Open Source Probabilistic Information Retrieval library. On line http://www.xapian.org [Visited 13/07/2005].

Appendix

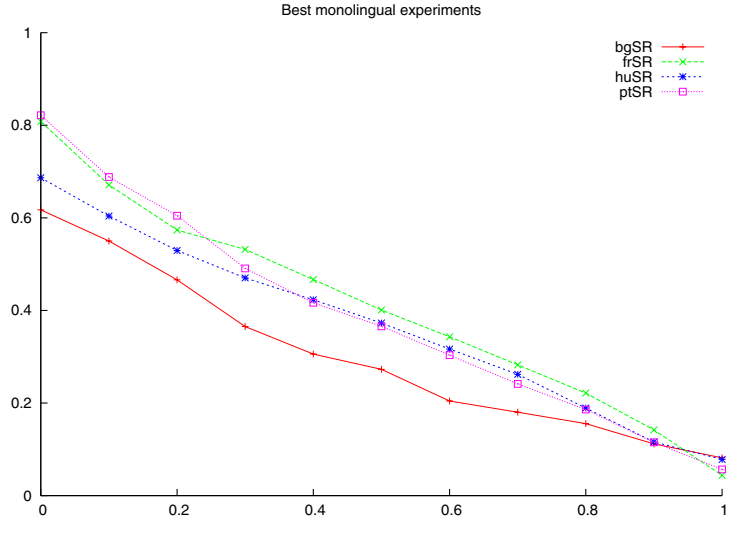

Fig. 1. Comparison of results from the best monolingual experiments

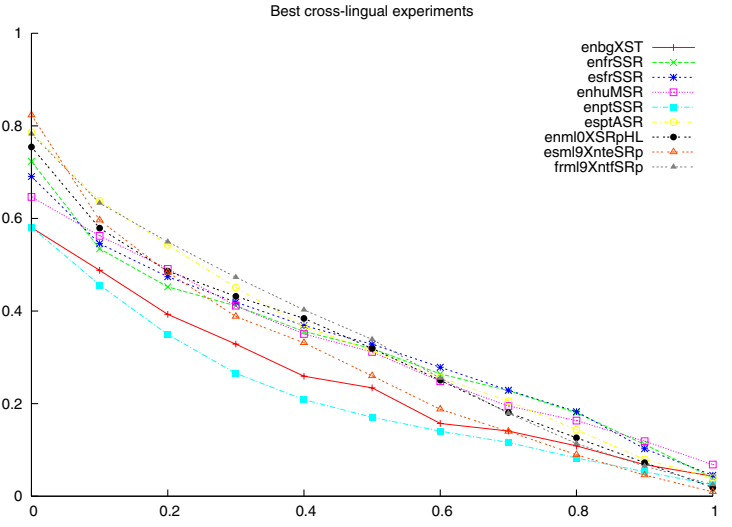

Fig. 2. Comparison of results from the best cross-lingual experiments