MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research

José L. Martínez[1], Julio Villena[2,3], Jorge Fombella[3], Ana G. Serrano[4],
Paloma Martínez[1], José M. Goñi[5], and José C. González[3,5]

[1] Computer Science Department, Universidad Carlos III de Madrid,
Avda. Universidad 30, 28911 Leganés, Madrid, Spain
{pmf,jlmferna}@inf.uc3m.es
[2] Department of Telematic Engineering, Universidad Carlos III de Madrid,
Avda. Universidad 30, 28911 Leganés, Madrid, Spain
jvillena@it.uc3m.es
[3] DAEDALUS – Data, Decisiond and Language,
S.A. Centro de Empresas "La Arboleda", Ctra. N-III km. 7,300 Madrid 28031, Spain
{jvillena,jfombella,jgonzalez}@daedalus.es
[4]Artificial Intelligence Department, Universidad Politécnica de Madrid,
Campus de Montegancedo s/n, Boadilla del Monte 28660, Spain
{agarcia,aruiz}@isys.dia.fi.upm.es
[5] E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid,
Avda. Ciudad Universitaria s/n, 28040 Madrid, Spain
jmg@mat.upm.es

Abstract. This paper describes the first set of experiments defined by the MIRACLE (Multilingual Information RetrievAl for the CLEf campaign) research group for some of the cross language tasks defined by CLEF. These experiments combine different basic techniques, linguistic-oriented and statistic-oriented, to be applied to the indexing and retrieval processes.

1 Introduction

It is well known that the amount of Internet pages is expanding rapidly; more and more encyclopaedia, newspapers and specialised sites related to almost every topic appear on-line and this has brought about the development and commercialization of a variety of tools devoted to facilitating information location and extraction from the billions of pages that make up the web. Among these tools we can find famous web search engines such as Google, Yahoo!, Altavista, etc. The need to process this huge amount of data has lead to important innovations in the field of Information Retrieval, most of them implemented into the aforementioned web search engines. Moreover, information is not only present in different kinds of formats but also in almost all languages used around the world.

There are currently three main trends in the field of the characterization of documents and queries which affect the information retrieval process: *semantic*

C. Peters et al. (Eds.): CLEF 2003, LNCS 3237, pp. 210–219, 2004.

approaches try to implement some degree of syntactic and semantic analysis of queries and documents, reproducing in a certain way the understanding of the natural language text; *statistical approaches* retrieve and rank documents according to the match of documents-query in terms of statistical measures and *mixed approaches* that combine both of them, trying to complement the statistical approach with semantic approaches by integrating natural language processing (NLP) techniques, in order to enhance the representation of queries and documents and, consequently, to produce adequate levels of recall and precision. Of course, there are other proposals concerning the Semantic Web that include a new layer on top of the search systems which is in charge of extracting information from web pages. Although the Semantic Web promises to be the future of text search systems, the work presented in this paper does not include this information representation subsystem.

The MIRACLE approach focuses on the mixed approach dealing with a combination of statistical and linguistic resources to enable the multilingual search to be carried out.

2 System Architecture

Several, free distribution and proprietary, components have been used to built the system architecture. These components are:

- *Retrieval Engine:* The information retrieval engine at the base of the system is the Xapian system [9]. This engine is based on the probabilistic retrieval model and includes a variety of functionality, useful for experiment definitions, e.g., stemmers based on the Porter algorithm [11].
- *Linguistic Resources:* Stemmers based on the Porter algorithm, included in the Xapian engine have been applied. Ad hoc tokenizers have also been developed for each language, standard stopword lists have been used and a special word decompounding module for German has been applied. Using EuroWordNet [10] to apply semantic query and index term expansions was not considered due to previous results obtained in CLEF campaigns. Retrieval precision fell to very low values.
- *Translation Tools:* For translation purposes, several different translation tools have been considered: Free Translation [6], for full text translations, LangToLang [7] and ERGANE [8], for word by word translations. Other available tools such as Google Language Tools [4] and Altavista Babel Fish [5], were tested but discarded.

The modular approach followed to build the architecture has provided the necessary flexibility and scalability to carry out the different defined experiments.

3 Experiment Definition

As is already known, Multilingual Information Retrieval (MIR) is the task of searching for relevant documents in a collection of documents in more than one language in response to a query, and presenting a unified ranked list of documents regardless of the language. Multilingual retrieval is an extension of bilingual retrieval, where the collection consists of documents in a single language that is different from the query language.

We have taken a number of factors which can dramatically influence system performance, into account when building our MIR system:

Combination Operator: As previously mentioned, our system is based on a probabilistic retrieval model, where several Boolean operators can be applied to construct a query. These are basically ‘AND’ operators and ‘OR’ operators, with the ability to assign weights to each operator. Another kind of operator investigated consists in the representation of the query as a document, indexing this new document and using acquired weights to build a new query. This operator is denoted with the suffix *doc* in our experiments, and tries to resemble a Vector Space Model Approach [3].

Stemming Algorithm: The stemming process is used to group together all words with related meanings under the same canonical representative. This grouping is guided by syntactical information, since words are arranged according to their stems. This dimension is used to take into account the effect of this stemming process on the use of original words to build the query. Of course, quality related to the stemming process is also relevant for system performance.

Techniques to Merge Retrieval Results: MIR systems are commonly based on three different approaches: the first translates the query into each target language and uses each translated query to search the independent collections according to the document language; in the second, all documents are translated to the language used to formulate the query, matching the query against the translated collection; in the third approach, the query is again translated to each target language, but all translations and the original query are used to build a multilingual one, which is applied to a unique document collection made up of documents in all languages. The MIRACLE contribution has taken into account the first and third approaches, but not the second due to the excessive resources needed to translate all the documents. With the first approach, techniques to merge the separate results lists obtained are needed. Techniques considered were:

- Round Robin, where results are merged taking into account positions in the results lists obtained for each language. So, if there are four target languages, the first element of each list is taken to obtain the four initial positions of the final results list, and so on.
- Normalization, where partial similarity measures are normalized (taking into account the number of documents in each collection) and ordered according to this normalized relevance value.

Of course, this dimension has no effect when monolingual or bilingual tasks are considered.

Translation Tools Used: Several on-line translation tools were considered for the experiments carried out by the MIRACLE team in the CLEF forum. These tools were:

Free Translation, for full text translations, LangToLang and ERGANE, for word by word translations. Different experiments have been defined according to the number of translation tools used. It is worth mentioning that retaining ambiguity often has a positive effect on MIR systems; in monolingual information retrieval there are several studies showing that dealing with lexical variation (discriminating word senses) is more beneficial for incomplete and relatively short queries, [2], due to the retrieval process itself carrying out a disambiguation process in extended queries (it is expected that a conjunction of terms would eliminate many of the spurious forms). Obviously, this dimension is not considered for monolingual experiments.

Query Section: As described in [14] queries are structured into three different fields: title, description and narrative. According to the query sections used, different experiments have been carried out, trying to take into account the relevance in performance introduced by long queries.

Relevance Knowledge: To improve the quality of retrieval results, knowledge on relevance of documents (supplied by the user) for a first query execution can be exploited. So, retrieved relevant documents can be used to remake the query expression and search again. The automatic relevance feedback process implemented consists of formulating a query, getting the first 25 documents, extracting the 250 most important terms for those documents, and constructing a new query to be carried out against the index database.

Tables 1, 2 and 3 show the different experiments submitted to CLEF 2003 for each task. Some details of these experiments should be commented:

- The *Tordirect* multilingual test applies the third approach described for MIR systems: the original query and its translations are used to build a query that is executed against a single index of all documents, regardless of the language.
- The *Tor3full* bilingual experiment includes the query in its original language to take into account the effect of erroneous translations.

Table 1. Monolingual Experiments

Exp. Identifier	**Combination Operator**	**Stemming Applied**	**Query Section Used**	**Rel. Feed back**
or (B)	OR	Yes	Title + Desc.	No
orand	AND for most frequent query stems, OR for the rest	Yes	Title + Desc.	No
Doc	DOC	Yes	Title + Desc.	No
Orfull	OR	Yes	Title + Desc. + Narr.	No
Orlem	OR	Yes + original query words	Title + Desc.	No
Orrf	OR	Yes	Title + Desc.	Yes

Table 2. Bilingual Experiments

Exp. Identifier	Combination Operator	Stemming Applied	Translators Used	Query Section Used	Rel. Feed back
Tor1 (B)	OR	Yes	FreeTranslation	Title + Desc.	No
Tor2	OR	Yes	FreeTranslation + LangToLang	Title + Desc.	No
Tor3	OR	Yes	FreeTranslation + Ergane	Title + Desc.	No
Tdoc	DOC	Yes	FreeTranslation	Title + Desc.	No
Tor3full	OR + original query words	Yes	FreeTranslation + Ergane	Title + Desc.	No

Table 3. Multilingual Experiments

Exp. Identifier	Combination Operator	Stemming Applied	Results Mixing Method	Translators Used	Query Section Used
Torall (B)	OR	Yes	Normalize	FreeTranslation	Title + Desc.
Torallrr	OR	Yes	Round Robin	FreeTranslation	Title + Desc.
Tor3	OR	Yes	Normalize	FreeTranslation +Ergane	Title + Desc.
Tdoc	DOC	Yes	Normalize	FreeTranslation	Title + Desc.
Tordirect	OR + original query words	Yes	Unique Index Database	FreeTranslation	Title + Desc.

4 Tasks and Results

This section contains the results obtained for tasks in which the MIRACLE consortium took part.

4.1 Multilingual-4

The languages selected by the MIRACLE research team were: Spanish, English, German and French. Four different experiments, all with Spanish as the query language, were carried out for this task, corresponding to those defined in the previous section.

Figure 1 shows Recall – Precision values obtained for each experiment. Best results correspond to *Tordirect*, followed by *Tor3*. Thus, we obtained better results when there is only one index database in which all languages are included. This can be due to variations in frequency of appearance of words that remain in the same form independently of the language considered, such us proper nouns.

The worst results were obtained by the approach where the retrieved documents list is put together taking into account the order of the results in the partial results list, i.e., when a round robin mixing technique is applied. This is not surprising taking into account that no method for considering the document collection as a whole to weight results is being applied.

If the values for average precision for all submissions are considered, the results of the MIRACLE approach are far from the best averages obtained. The baseline for our multilingual tasks was *Torall*, which has been improved by the *Tordirect*, *Tor3* and *Tdoc* experiments. Some conclusions could be drawn:

- The third approach for multilingual processing, where a single multilingual index is built for the document collections and for the query, could lead to better results than separately indexing the collections.
- If several translation tools are applied, precision can be improved, perhaps due to the inclusion of a great variety of translations for each query word.
- The *doc* technique can offer better results because the representation built for the query is closer to the document representations (remember that the query is indexed as part of the document collection)
- Our worst result was obtained for the *QTorallrr* experiment, due to the method applied to merge partial results list when constructing the final retrieved documents list.

4.2 Bilingual

For the bilingual task, three different language combinations were used: Spanish queries against the English document collection, French queries against the English document collection and Italian queries against the Spanish document collection. The experiments carried out for each language pair were very similar to those described for the previous task. This is the first time that the MIRACLE research team takes part in CLEF, so it was possible to choose English as one of the target languages for this task.

Figure 2 shows the results for each bilingual task. For technical reasons it was not possible to run *Tor1* or *Tor3* for the bilingual French – Spanish task. As the graphics show, the best results for all language combinations are obtained for *Tor1*. This result seems to show that using several translation tools does not improve the results, which appears to be inconsistent with the conclusions drawn from our multilingual experiments. The explanation is that, for multilingual experiments, three translations for each query are obtained and used to construct the query, which can lead to a more complete query representation, but for bilingual experiments, only one translation is obtained and only one document collection is searched. The narrative field for queries offers the worst retrieval performance, perhaps due to the excessive number of terms introduced when considering all query fields and translating them with all available tools.

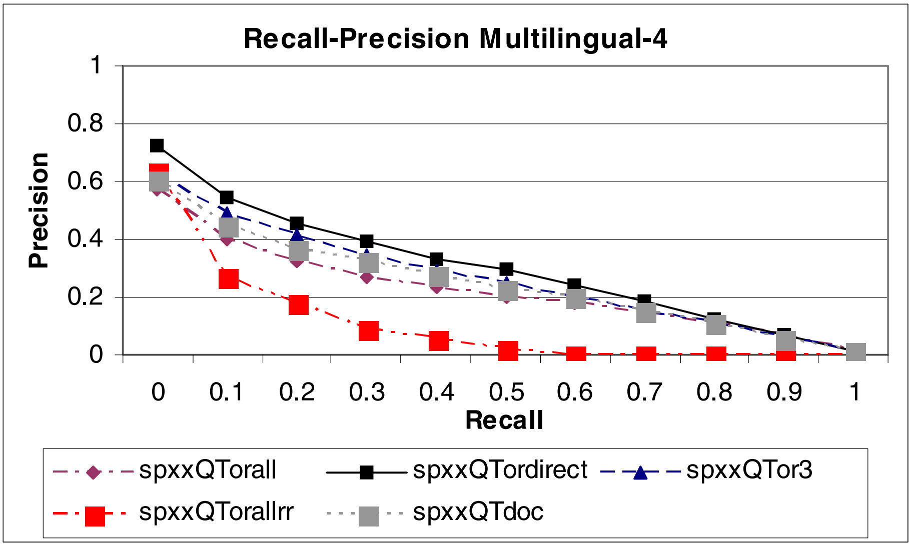

Fig. 1. Recall-Precision graph for the Multilingual-4 task

A comparison with the result of the rest of the participants in CLEF 2003 has been made using mean average precision values supplied with the result files. The results for the Italian–Spanish tasks are not as good as the rest of the submissions. Our best system is performing below the best of all submissions as is our mean precision value. Of course, it must be taken into account that our results are included in average precision values provided by the organisation. On the other hand, for the Spanish–English and the French–English tasks, the performance obtained is the best of all of the participants in this task.

4.3 Monolingual

In this task only one language was used to formulate queries which are processed against a document collection in the same language as the query. The MIRACLE research team submitted runs for the Spanish, English, French and German tasks.

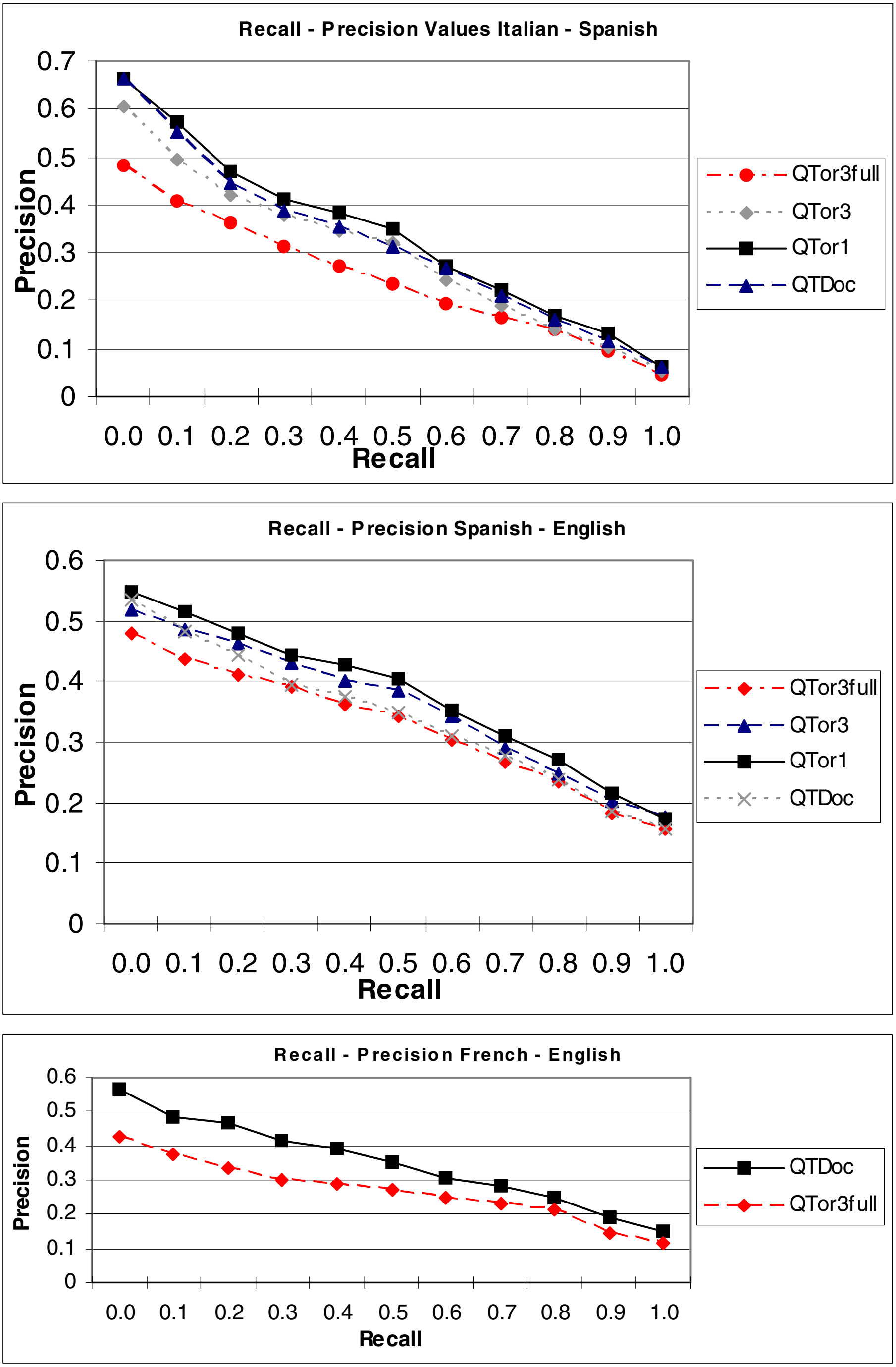

Fig. 2. Recall-precision graphs for bilingual tasks

Several different experiments were carried out for this task, as described in Table 1. Taking into account obtained results, only for the French – French task have we improved on the baseline experiment, consisting of an ORed expression made up of all words in the query. For the rest of the tasks, variations in the baseline experiment have not lead to better Recall- Precision values. Tasks where relevance feedback has been used always give the worst results, suggesting that our relevance feedback method should be changed. CLEF 2003 participants who applied relevance feedback improved their retrieval results. Experiments where the query is used to construct a document to be indexed and used as a query expression to be matched against the index document database, always resulted in lower performance values than the baseline. Again, this fact seems inconsistent with multilingual conclusions, but for this experiment only one language is being considered and, probably, the *doc* method has to be adapted for this particular case.

To compare MIRACLE results with all participants in CLEF, average precision values provided by the CLEF organisation are used. MIRACLE monolingual French – French results lead to low precision values. This can be due to the linguistic resources used for this language, e.g., the tokenizer used is not specific for the French language, producing low quality stems. Also, the French – French task is the only one where the best of our runs does not reach the mean value for all runs submitted. In the German – German task, results are not much better, maybe for a similar reason.

5 Conclusions and Future Directions

As a first conclusion from the experiments carried out, none of the different techniques applied improves results obtained for defined baseline experiments. Although the MIRACLE approach has obtained good results for bilingual tasks working on the English collection, the MIRACLE results do not improve the retrieval performance achieved by the best participants in the CLEF 2003 initiative. Nevertheless, the objectives of this research team have been accomplished. The main goal pursued with this first participation in the CLEF initiative was to establish a starting point for future research work in the field of cross-language retrieval. For later CLEF initiatives, according to results obtained, new experiments will be defined, aimed at looking deeply into the proposed mixed approach. Improvements will apply different retrieval models, in particular, the Vector Space Model, supported by a semantic approach, and will follow two basic lines ([1],[12],[13]):

- From the linguistic point of view, specific linguistic resources and techniques will be applied, such as shallow parsers, tokenizers, language specific entity recognition subsystems and semantic information, probably extracted from EuroWordnet.
- From the statistical perspective, ngram approaches will be implemented. Some of the CLEF 2003 participants have obtained good results with ngram techniques and the MIRACLE team will try to improve on these results combining some of the above mentioned linguistic techniques. Several weight assignment methods will also be explored.

Acknowledgements

This work has been partially supported by the projects OmniPaper (European Union, 5th Framework Programme for Research and Technological Development, IST-2001-32174) and MIRACLE (Regional Government of Madrid, Regional Plan for Research, 07T/0055/2003).

References

[1] Greengrass, E. Information Retrieval: A Survey, Internet Available (20.10.2003): http://www.csee.umbc.edu/cadip/readings/IR.report.120600.book.pdf, November (2000).

[2] Voorhees, E.: On expanding query vectors with lexically related words, 2nd Text Retrieval Conference, pp. 223-231, (1994).

[3] Karen Sparck Jones and Peter Willet: Readings in Information Retrieval, Morgan Kaufmann Publishers, Inc. San Francisco, California (1997).

[4] "Google Language Tools", www.google.com/language_tools.

[5] "Altavista's Babel Fish Translation Service", www.altavista.com.

[6] "Free Translation", www.freetranslation.com.

[7] "From Language To Language", www.langtolang.com.

[8] "Ergane Translation Dictionaries", http://dictionaries.travlang.com.

[9] "The Xapian Project", www.sourceforge.net.

[10] "Eurowordnet: Building a Multilingual Database with Wordnets for several European Languages." http://www.let.uva.nl/ewn/, March (1996).

[11] "The Porter Stemming Algorithm" page maintained by Martin Porter. www.tartarus.org/~martin/PorterStemmer/.

[12] Sparck Jones, K. Index term weighting. Informa. Storage and Retrieval, 9, 619-633, (1973).

[13] Salton, G., Yang,C. On the specification of term values in automatic indexing. Journal of Documentation, 29 (1973), 351-372.

[14] Braschler, M. and Peters, C. CLEF2003: Methodology and Metrics, this volume.