Evaluation of MIRACLE approach results for CLEF 2003[1]

José Luis Martínez[1], Julio Villena Román[2,3], Jorge Fombella[3], Ana G. Serrano[4], Alberto Ruiz[4], Paloma Martínez[1], José M. Goñi[5], José C. González[3]

[1] Advanced Databases Group, Computer Science Department,
Universidad Carlos III de Madrid, Avda. Universidad 30,
28911 Leganés, Madrid, Spain
{pmf,jlmferna}@inf.uc3m.es, jvillena@it.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] ISYS group, Artificial Intelligence Department, Technical University of Madrid
Campus de Montegancedo s/n, Boadilla del Monte 28660, Spain
{agarcia,aruiz}@isys.dia.fi.upm.es

[5] Department of Mathematics Applied to Information Techmologies,
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 MIRACLE (Multilingual Information RetrievAl for the CLEf campaign) approach and results for the mono, bi and multilingual Cross Language Evaluation Forum tasks. The approach is based on the combination of linguistic and statistic techniques to perform indexing and retrieval tasks.

1 Introduction

It is well known that the number of Internet pages is expanding rapidly; more and more encyclopaedia, newspapers and specialised sites about almost every topic appear on-line and it has produced development and commercialization of a variety of tools devoted to facilitate information location and extraction from the billions of pages that conform the web. Among these tools we can find famous web search engines like Google, Yahoo!, Altavista, etc. The need of processing all this great amount of data has lead to important innovations in the field of information retrieval, most of them implemented into mentioned web search engines. Moreover, information is not only present in different kinds of formats but also in almost all languages used around the world.

Currently, there are three main trends in the field of characterization of documents and queries which affect the information retrieval process: *semantic 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 some statistical measure 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 over the search systems that is in charge of extracting information from web pages. Although Semantic Web promises to be the future of text search systems, the work presented in this paper does not include such information representation subsystem.

[1] Part of this work has been supported by OmniPaper (IST-2001-32174) and CAM 07T/0055/2003 projects

The MIRACLE approach is focused in the mixed approach dealing with a combination of statistical and linguistic resources to enable the multilingual search.

2 Three approaches to multilinguality

Multilingual Information Retrieval (MIR) shares many of the characteristics of the general IR problem. The classic IR paradigm reflects the desire of a user to get documents (abstracts, news, articles, web pages, books, or even images, audio clips or video fragments) about a certain topic. The user provides a free form textual description as a query and the IR engine derives a set of index terms, which then are matched against the index terms characterizing the whole collection. Finally, the documents that match best are returned to the user in a ranked order.

Precisely, 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 an 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.

There are three approaches to manage MIR tasks. The first approach translates the source topics separately into all the document languages in the document collection; then, monolingual retrieval is carried out separately for each document language, resulting in one ranked list of documents for each document language; finally, the intermediate ranked lists of retrieved documents, one for each language, are merged to yield a combined ranked list of documents regardless of the language. The second approach translates a multilingual document collection into the topic language; then, topics are used to search against the translated document collection. The third one also translates topics to all document languages as in the first approach. The difference is that the source topics and the translated topics are concatenated to form a set of multilingual topics. The multilingual topics are then searched directly against the multilingual document collection, which produces a ranked list of documents in all languages.

The later two approaches do not involve merging two or more ranked lists of documents, one for each document language, to form a combined ranked list of documents in all document languages. The merging task is hard and challenging. Up to date, no effective technique has been developed to face this matter.

The most extended approach is the first one. Translating large collections of documents in multiple languages into topic languages requires the availability of machine translation systems that support all the necessary language pairs, which is sometimes a problematic task. For example, if the document collection consists of documents in English, French, German, Spanish, Portuguese and Catalan, and the topics are in English, in order to perform the multilingual retrieval task using English topics, one would have to translate the French, German, Spanish, Portuguese and Catalan documents into English. There exist translators, such as Google Translate [4] or Babelfish [5], that can do the job for all languages except for Catalan. The availability of translation resources and the need for extensive computation are factors that limit the applicability of the second approach. The third approach is appealing as it does not require translating the documents, and circumvents the difficult merging problem. However, there is some empirical evidence showing that the third approach is less effective than the first one.

2.1 Evaluation of MIR systems

Classical IR measures of success are precision and recall but since MIR is in the intersection of IR and Machine Translation (MT), there are some specific problems that are not shared by traditional monolingual text retrieval.

Traditional IR works directly with the words used in the initial user query, and most of the effectiveness of IR systems comes from the matches among these query words (including morphological changes, as stemming) and the same words appearing in the relevant documents. The basic intuition is that as a document contains more words appearing in the user query, more likely the document is relevant to that query. In the case of MIR, the initial query is in one language and the documents are in another, so simply word matching mechanisms will rarely work (except pure coincidences and some proper names).

Three of the core components of MIR (more precisely, the first approach) are monolingual retrieval, topic translation and merging. Performing multilingual retrieval requires many language resources such as stopwords, stemmers, bilingual dictionaries, machine translation systems, parallel of comparable corpora, etc. The final performance of multilingual retrieval can be affected by many factors, such as monolingual retrieval performance of the document ranking algorithm, the quality and coverage of the translation resources, the availability of language-dependent stemmers and stopwords, and the effectiveness of merging algorithm. Since merging of ranked lists of documents is a challenging task, it is better to try to improve multilingual retrieval performance by improving monolingual retrieval performance and exploiting translation resources. Thus, the

first problem that a MIR system must solve is to use a good monolingual retrieval engine which offers the best possible precision/recall curve.
Then, the effectiveness of the translation has to be carefully taken into account, for which there exist three topics to study. The first one is to know how a term expressed in one language might be written in another. The second topic is deciding which of the possible translations is the best one. The third topic is deciding how to properly weight the different translation alternatives where more than one is retained.
Retaining ambiguity is often useful to improve MIR systems; in monolingual IR there are several studies showing that dealing with lexical variation (discriminating word senses in extended queries) is more beneficial for incomplete and relatively short queries, [2], due to the retrieval process itself performs a disambiguation process (it was expected that a conjunction of terms will eliminate many of the spurious forms). Consider the following example, the Spanish word “sal” can be translated into English as the noun “salt” or the imperative “go out”. In a translation process, the MT system should choose only one valid translation. However, in an IR process, it could be interesting to preserve both alternatives. If the original Spanish query is about “salt and diet” and the MIR system retains both “salt” and “go out”, then some noise may be introduced, but documents about “salt and diet” will be found. A translation system that chooses only one translation but the erroneous one, will not find any result.
The topic related to how treating alternatives is very important. For example, a document that contains one translation of each query term (in a query with two terms) would probably be more relevant than a document that contains many variants of the first term but none of the second.
The easiest way to find translations is to search for them in a bilingual dictionary, but this is not as easy as it may seem. Samples of problems are:

- Missing word forms: for example, there exists an entry for “electrostatic” but not for “electrostatically”, because a human reader can easily infer the meaning. Stemming entries could be a possibility, at expense of adding noise
- Spelling norms: only one form appears (e.g. colour or color)
- Spelling conventions: the use of hyphenation (e.g. fallout, fall-out, fall out)
- Coverage: general language dictionaries contain the most common words in a language, but rarer technical words are often missing.

3 MIRACLE experiments description

This section contains a description of the tools, techniques and experiments used by the MIRACLE team for the cross-lingual, monolingual and multilingual tasks.

The information retrieval engine in the base of the system was 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 Porter algorithm [11]. Also, ad hoc tokenizers have been developed for each language, standard stopwords 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.

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.

All these tools have been coupled to define the set of experiments. For the multilingual task, different methods to run queries and merging results have been applied, according to the different multilingual approaches described in section 2:

Torall: Query definition is translated to all target languages using FreeTranslation and independently executed against index databases for each language. Results are merged taking into account normalized relevance values obtained for each language. This experiment constitutes the baseline for the multilingual task.
Torallrr: Query execution is made as in the previous experiment, but results are merged taking into account positions in the results lists obtained for each language. So, if there are 4 target languages, the first element of each list are taken to obtain the four initial positions of the final results list, and so on.
Tor3: The query used is built using FreeTranslation and Ergane to translate the query to each target language. Then, query terms are ORed and executed against each language document database. Final result list is obtained normalizing relevance values for results lists retrieved for each language.

Tdoc: A new document is built using the original query and the translations to each target language. Translations are obtained using FreeTranslation. Then, a new document is built and used to query the different index databases. Results are merged using the normalized approach described in previous experiments.
Tordirect: Finally, a special feature of Xapian retrieval engine has been used to run queries against the indexed document database. This feature consists in the construction of a unique index database from the index databases built for each language. The query is built concatenating the original query and the translations obtained to each target language.

For bilingual tasks, several translation tools have been used, allowing the creation of the following experiments:

Tor1: Where the FreeTranslation tool has been used to translate the complete query, combining then stems with the OR operator. This is the baseline experiment for the bilingual tasks.
Tor2: In this case, two different translation tools are applied, FreeTranslation and LangToLang, combining query stems with the OR operator. This approach tries to produce better retrieval performance by adding different translations for the words in the query.
Tor3: Similar to the previous one, substituting LangToLang translator by Ergane. There were attempts including all the three translators but LangToLang was excluded due to low precision results.
Tdoc: Query definition is translated using FreeTranslation and the document obtained is used to query the document database. Again, an approximation to Vector Space Model is used, like in experiment identified as **doc**.
Tor3full: Defined as Tor3 but adding the query text in the original language, so query terms incorrectly translated or that does not have proper translation to other languages are included in their original form.

The experiments included in the monolingual task are:

or: This is the baseline experiment defined. The following ones will make some variations over this baseline, trying to discover benefits or drawbacks of the application of different statistical or linguistic approaches. So, this first monolingual experiment consists on the combination of stemmed words appearing in the title and description fields of the query using an OR operator.
orand: This experiment uses the same ORed combination of the previous one but using AND operator for stems that appear more frequently in the query definition. This approach tries to include statistical information about terms appearing in the query, increasing relevance of terms with greater frequency.
doc: For this experiment, a special feature of Xapian system is used, which allows execution of queries based on documents against the indexed document collection. This approach is similar to the application of the Vector Space Model.
orfull: Using the same ORed combination of experiment **or** but including stems appearing in the narrative query field too. So, all the information available for the query expression is used, trying to use it to improve retrieval performance.
orlem: Joins original words (without stemming) and stems into a single query using the OR operator to concatenate query terms. This experiment tries to measure the effect of inadequate proper noun translations by adding the original form of the query terms.
orrf: Includes some kind of relevance feedback, based on results for previous queries. The process consists on executing a query, getting the first 25 documents, extracting the 250 most important terms for those documents, and building a new query to execute against the index database.

4 Tasks submitted and obtained results

Once the different experiments have been described, this section contains results obtained for tasks where the MIRACLE consortium has taken part in.

4.1.1 Multilingual-4

Languages selected by MIRACLE's research team have been: Spanish, English, German and French. Four different experiments, all of them taking Spanish as the query language, have been defined for this task, corresponding to the ones defined in the previous section.
Figure 1 shows Recall – Precision values obtained for each experiment. Best results correspond to **Tordirect**, followed by **Tor3**. So, better results are produced when there is only one index database where all languages are

included. This can be due to variations in frequency of appearance of words that keep in the same form independently of the language considered, such us proper nouns.
The worst results are produced by the approach where the retrieved documents list is built taken into account the order of results in partial results list. This is not surprising taking into account that no method to consider the document collection as a whole to weight results is being applied.

If values for average precision for all submissions are considered, results of the MIRACLE approach are far from the best average obtained for all submissions. On the other hand, median values for average precision are in the range defined by the rest of the submissions. The best MIRACLE run is over the median value provided when all submissions are considered (including our submission). The baseline for our multilingual tasks was **Torall**, which has been surpassed by the **Tordirect**, **Tor3** and **Tdoc** experiments. The worst result has been obtained for the QTorallrr experiment, due to the method applied to merge partial results list for building the final retrieved documents list.

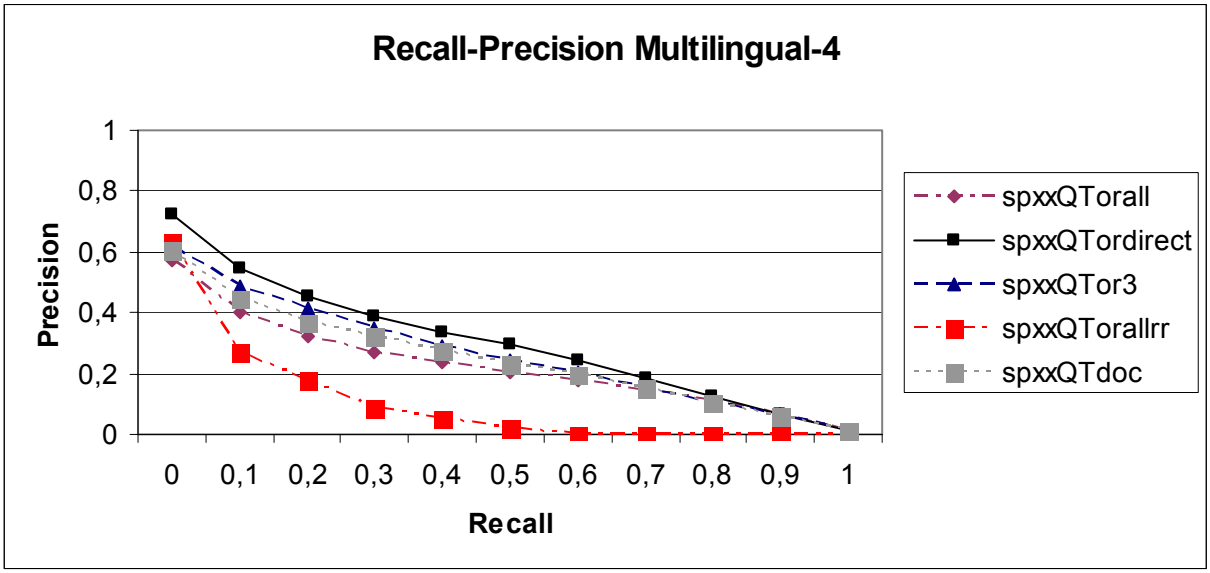

Figure 1. Recall-Precision graph for the Multilingual-4 task

4.1.2 Bilingual

For the bilingual task, three different language combinations have been defined: Spanish query against the English document collection, French query against the English document collection and Italian query against the Spanish document collection. Experiments defined for each language pair have been very similar to the ones described for the previous task. This is the first attempt for MIRACLE research team to take part in CLEF, so it has been possible to choose English language as one of the target languages for this task.

Figure 2 shows results for each bilingual task. Due to technical reasons, was not possible to run **Tor1** nor **Tor3** for the bilingual French – Spanish task. As graphics show, best results for all languages combinations are obtained for **Tor1**, meaning that using several translation tools does not improve results. Using 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.

Comparison with the rest of participants at CLEF 2003 has been done using mean average precision values supplied with the result files. Figure 3 shows a graphical view of this comparison values. The Italian – Spanish tasks is far form the rest of submissions, our best effort is less than the best for all submissions as for 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 gap between the best precision values for all submissions and our best results is not so large, and average figures for our results are better than average figures for the rest of submissions.

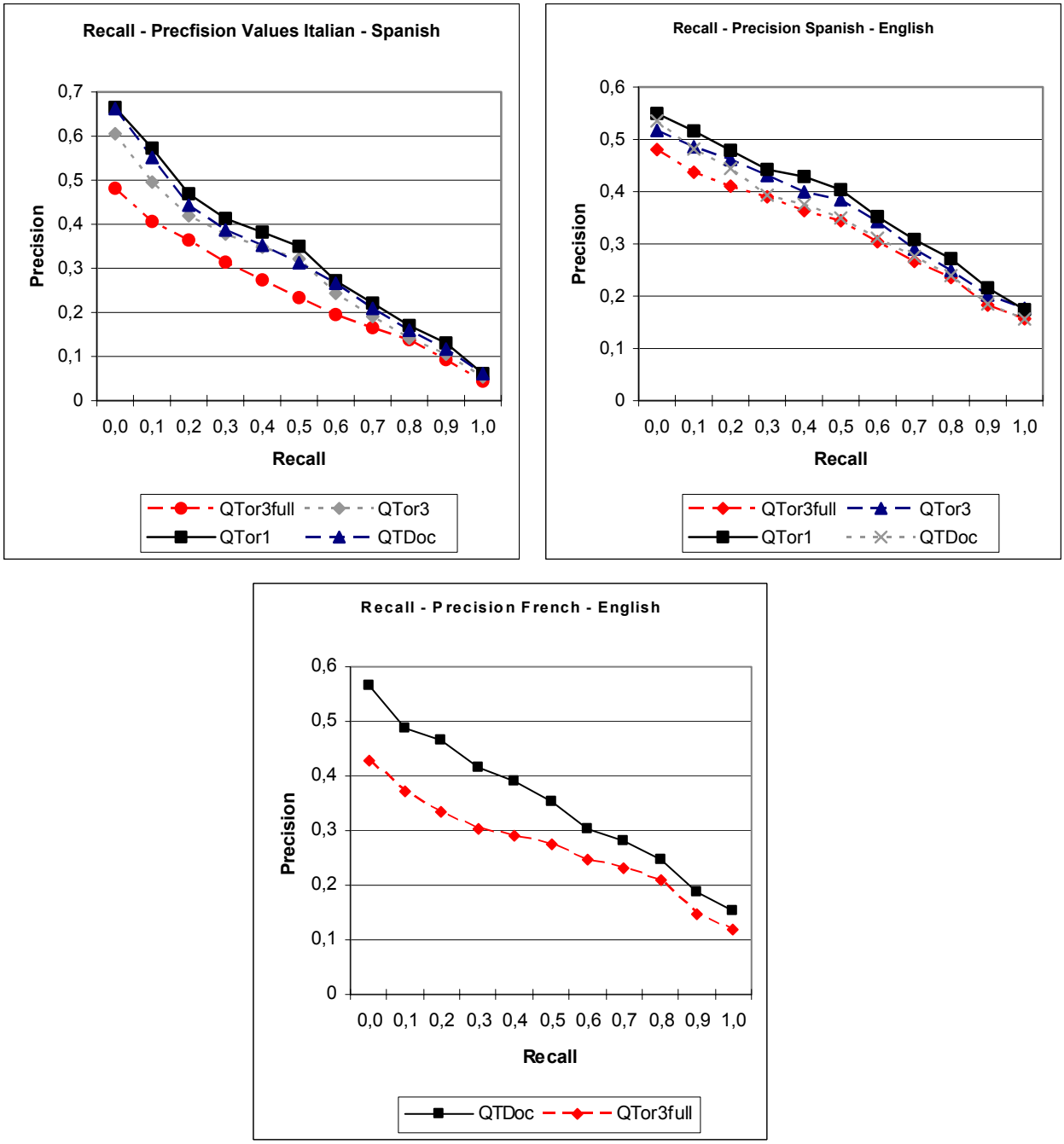

Figure 2. Recall-Precision graphs for bilingual tasks

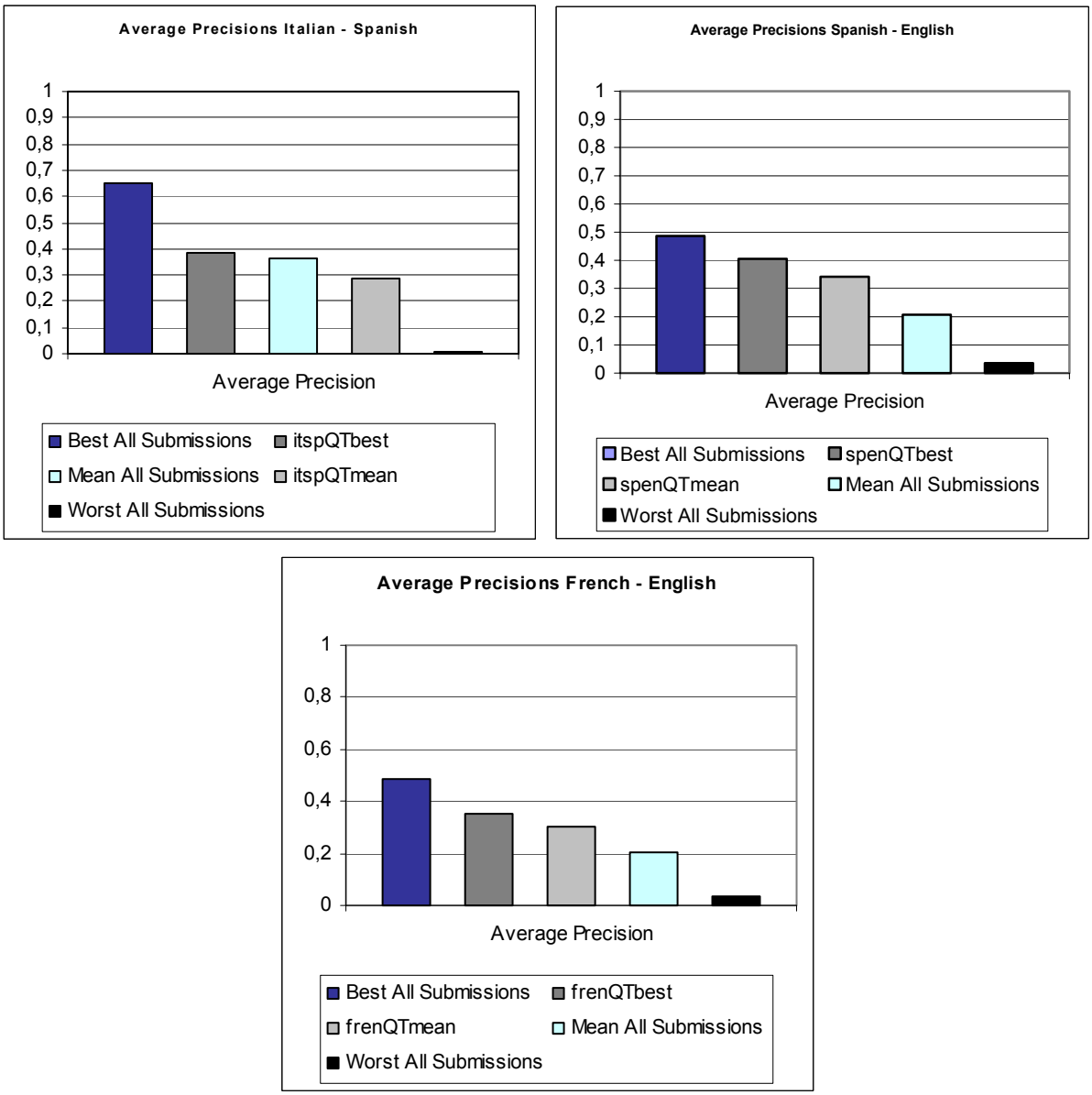

Figure 3. Precision comparison with all runs submitted for the bilingual task

4.1.3 Monolingual

In this task only one language will be used to express queries, which will be processed against a document collection in the same language as the query. MIRACLE's research team has submitted different runs for Spanish, English, French and German languages.
Several different experiments have been defined for this task, as defined in section 3. Recall-Precision values obtained for this set of experiments are shown in Figure 4.

As shown by this graphic representation, only for the French – French task the baseline experiment, consisting in an ORed expression formed by all words in the query, has been improved. For the rest of the tasks, variations of the baseline experiment have not lead to better Recall- Precision values. Tasks where relevance feedback has been used, provides always the worst results, so it is possible that the relevance feedback method have to be changed. Experiments where the query is used to build a document to be indexed and used as a query expression to be executed against the index document database, always produce lower performance values than the baseline.

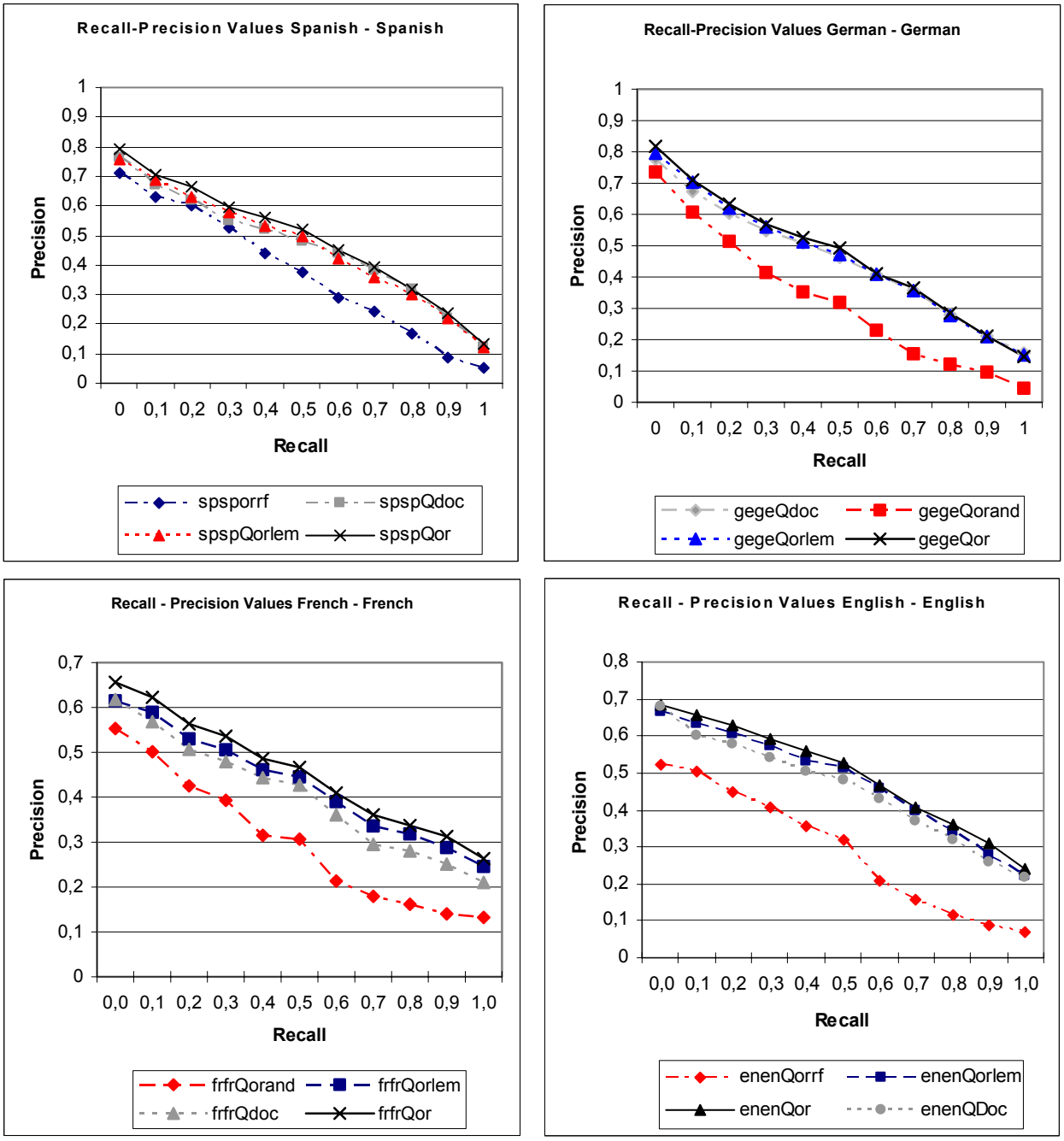

Figure 4. Recall-Precision graphs for the monolingual tasks

To compare MIRACLE results with all participants at CLEF, again average precision values provided by the CLEF organisation. Figure 5 shows compared average precision figures.

This graphical representation shows that, comparing with the rest of participants, MIRACLE monolingual French – French results lead to low precision values. This can be due to the linguistic resources used for this language, e.d., 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 due to a similar reason.

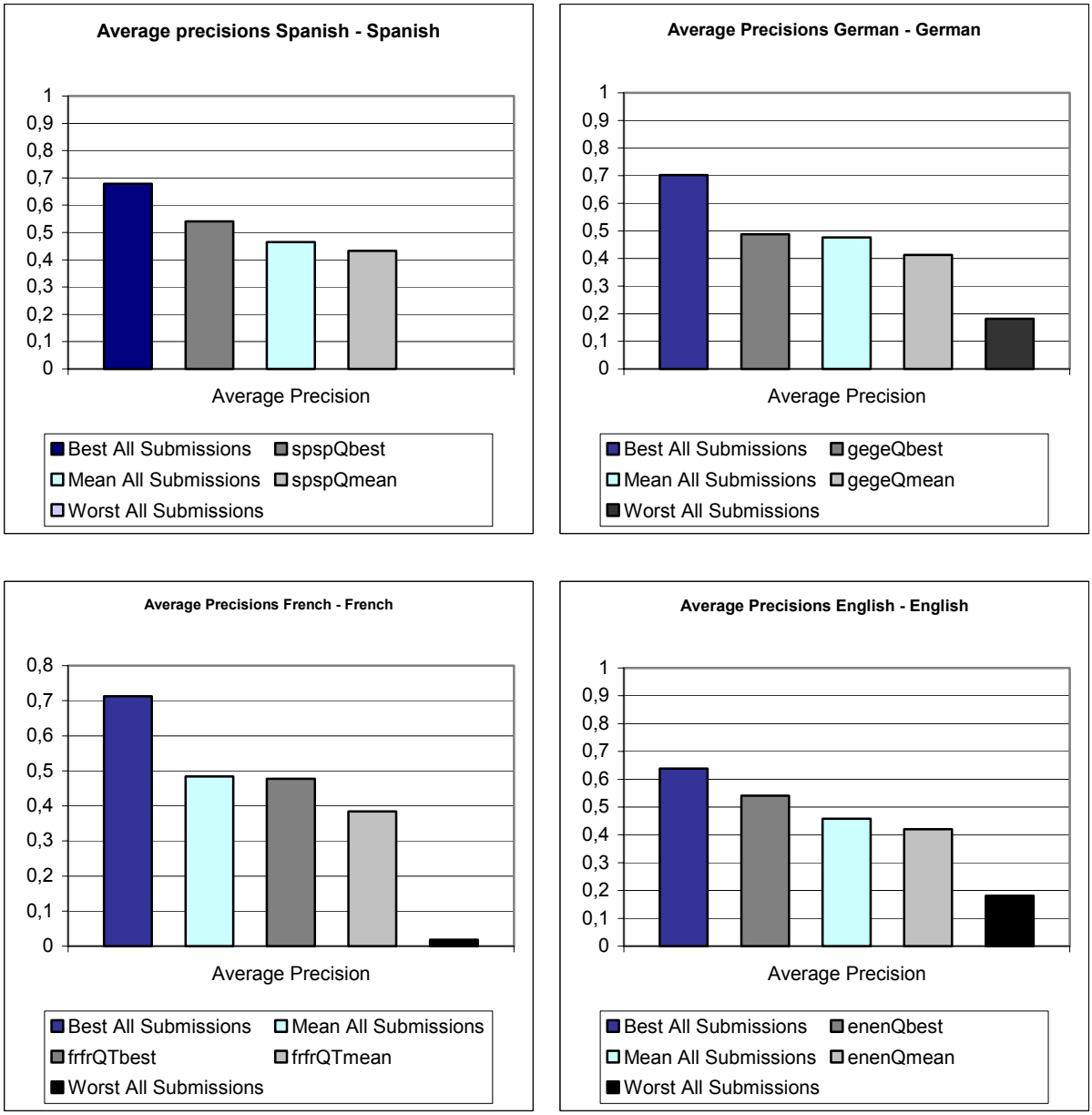

Figure 5. Precision comparison with all runs submitted for the monolingual task

5 Conclusions and Future Directions

The main conclusion that can be extracted taking into account obtained results is that no one of the different approaches studied improves results obtained for our baseline experiments. Although MIRACLE approach results are far from the best retrieval performance figures obtained by the rest of participants in the CLEF 2003 initiative, 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 cross-language retrieval field. Taking into account obtained results, new approaches will be defined to improve multilingual and monolingual retrieval performance. One possible line of work is to change the retrieval model used, using the Vector Space Model to index documents, supported by a semantic approach. Also, linguistic resources and techniques will be introduced, like specific parsers where more linguistic knowledge can be included. Also, combination with some statistical methods, like ngrams or different weight assignment methods, can be studied.

References

[1] Greengrass, E. Information Retrieval: A survey, November (2000).
[2] E. Voorhees, On expanding query vectors with lexically related words, 2nd Text Retrieval Conference, pp. 223-231, 1994.
[3] Karen Sparck Jones y 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", world.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/