MIRACLE evaluation of results for ImageCLEF 2003[1]

Julio Villena Román[2,3], José Luis Martínez[1], 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. ImageCLEF is a new pilot experiment introduced in CLEF 2003. It is devoted to the cross language retrieval of images using textual descriptions related to images contents. This paper presents MIRACLE research team experiments and results obtained for this track.

1 Introduction

There are several differences between CLIR (Cross Lingual Image Retrieval) for documents and images, due to the unlike nature of both information structures. Although documents and images can be thought as being similar, in a certain level of abstraction (in that both of them express ideas and concepts), practical applications that deal with them must take into account their differences.

The main difference between a document and an image is that the latter can usually be interpreted in several different ways, while a document could be, more or less, understood in a certain manner. This is one of the beauties of images and non-verbal communication, but also one of its big problems while trying to work with it in an ordered and structured manner.

During last years, great efforts have been made in the analysis and study of content-based image retrieval research; to the time of this writing, it is almost clear that, for an important period of time, these kind of techniques are not going to solve the problem. This is the main reason to focus on image retrieval based on text descriptions and keywords. The idea is to use a textual description of each image as the base for the image retrieval process. This approach has two main drawbacks:

- Image descriptions will be incomplete, as it happens for text documents.
- Image descriptions would be usually quite short, typically image captions and/or a few keywords referring the most relevant characteristics and components of the image.

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

- The multilingual dimension of the problem. Although images are not coupled to an specific language, image captions will be available in different languages, so some kind of multilingual approach must be considered.

But there are also some advantages related to the use of image descriptions in image retrieval applications. One of them is extensibility to other multimedia information formats, like audio or video, or even other kind of information, e.g., source code. On the other hand, image retrieval would be very interesting for appliance in online newspapers, reviews, television, etc.

Techniques applied by the MIRACLE research team to this task go from relevance feedback to topic term semantic expansion using WordNet. The main idea behind MIRACLE participation is to compare how these different retrieval techniques affect retrieval performance.

2 ImageCLEF track description

In order to experiment with image CLIR a collection of nearly 30,000 black and white images from the Eurovision St Andrews Photographic Collection where provided. Each image had a quite large English caption (of nearly 50 words). On the other hand, a set of 50 queries in English, French, German, Italian, Spanish and Norwegian was also provided. Non English queries have been obtained as a human translation of the original English queries, which also included a narrative explanation of what should be considered relevant for each image.

The tasks proposed were to retrieve the relevant images of the collection using different query languages. Therefore, this year, the ImageCLEF track only dealt with monolingual and bilingual image retrieval. Multilingual image retrieval is supposed to be close to multilingual document retrieval, both in techniques and expected results, and so it has not been considered this year. After all, a first logical step to deal with multilingual retrieval is trying to solve (up to a reasonable point) monolingual and bilingual retrieval problems.

Although there are clear limitations in current ImageCLEF track, both in the size of the collection and the number of possible experiments to perform (six –one monolingual and five bilingual), it is an interesting starting point to grasp an idea of how good (or bad) the performance of this kind of systems are, both in monolingual and bilingual searches.

For this task, the MIRACLE team has submitted 25 runs, 5 for the monolingual English task, 6 for the bilingual Spanish to English task, 6 for the bilingual German to English task, 4 for the bilingual French to English task and, finally, 4 for the bilingual Italian to English task. All tasks submitted are automatic tasks.

3 MIRACLE experiments description

This section contains a description of the tools, techniques and experiments used by the MIRACLE team for the different tasks attended for this ImageCLEF campaign.

As for the cross language tasks, the information retrieval engine used at the core of the system, has been Xapian [5], based on the probabilistic retrieval model. This tool has a high configuration level, allowing the use of different techniques related to information retrieval tasks, such as stemmers based on Porter algorithm [7].

In order to apply natural language processing to image descriptions and queries, ad hoc tokenizers have been developed for each language, allowing to recognize some of the usual compound words each language has, in addition to identify different kinds of alphanumerical tokens such as dates, proper nouns, acronyms, etc. Standard stopwords lists have also been used and a special word decompounding module for German queries has also been applied. WordNet [6] has been used to expand queries.

For translation purposes, two available translation tools have been considered: Free Translation [3] for full text translations, and ERGANE [4] for word by word translations.

These tools have been coupled in different ways, in order to evaluate different approaches and compare the influence of each one in the precision and recall of the image retrieval process.

In particular, the experiments submitted for the monolingual task have been the following:

OR: Intended as the baseline experiment, to compare with results of other expirements, consists on the combination of all the stemmed words appearing in the title of the query, without stopwords, using an OR operator between them.
ORlem: This experiment joins the original words of the query and its stems in a single query, using the OR operator to concatenate them. The idea behind this experiment is to measure the effect of inadequate stemming of words, by adding the original form of the query terms.
ORlemexp: The idea behind this experiment is to make synonym expansion of the terms and stems used in the ORlem experiment, linking the obtained words with an OR operator. The pretended result is to retrieve a larger number of documents (increase recall), despite the possible penalization in precision we could have.
Doc: For this experiment, a special feature of Xapian system is used, which allows execution of queries based on documents against the indexed document collections. This approach is similar to the application of the Vector Space Model. In order to carry out this experiment, the query is first indexed as if it was another image description, and then similar documents are retrieved.
ORrf: This experiment performs a blind relevance feedback (based on the results of a simple OR query). 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, which would provide the final results.

Bilingual experiments submitted have been the following:

TOR1: Similar to the monolingual OR experiment, but using the FreeTranslation tool to translate the complete query. Therefore, the steps followed to build the query are: first, translate the full query using FreeTranslation, then use the tokenizer to identify the different tokens in English, extract the stems of the tokens, remove stop words (in these case stop stems) and generate an OR query using the resulting terms.
TOR3: In this case, in addition to the translation of the complete query, a word by word translation is added, using ERGANE. The following steps (tokenizing, stemming and OR concatenation) are the same as TOR1 experiment. The idea is to improve retrieval performance by adding different translations for the words in the query.
Tdoc: This is the bilingual equivalent of the monolingual Doc experiment. This time the query is first translated using FreeTranslation and the result obtained is indexed in the system as if it were just another image description. The information retrieval engine (Xapian) is then asked to retrieve similar images to this newly added one.
TOR3exp: This is the bilingual equivalent of the monolingual ORlemexp experiment. It is basically the same as the TOR3 experiment, but adding a synonym expansion (using Wordnet) of the translated terms.
TOR3full: Similar to the TOR3 experiment, but adding the original query (in the original language) to the terms used in the OR query. This way, query terms incorrectly translated or that do not have a proper translation into English are included in their original form (possibly being of little interest, but at least appearing somehow).
TOR3fullexp: This experiment is a combination of TOR3full and TOR3exp, using both translation engines together with the original query, adding synonym expansion for all the terms obtained.

4 Tasks submitted and obtained results

In this section, results obtained by the MIRACLE team will be presented and compared, in order to infer some conclusions relative to the different approaches tested.

To assess the defined experiments, CLEF evaluation staff used the first 100 results of each submission (45 in total) to make a document pool (different for each query). In addition, the results of different interactive searches manually performed by assessors were also added to each pool. Then, two different assessors evaluated all the documents in the pools, considering a ternary scale: **relevant**, **partially relevant** and **not relevant**. The partially relevant judgment was used to pick up images where the judges thought were in some way relevant, but could not be entirely confident.

As a final step, four relevance sets were created using the relevance judgments of both judges:

Union-strict: The images of this set were the union of the ones judged as relevant by any assessor.
Union-relaxed: The images of this set were the union of the ones judged as relevant or partially relevant by any assessor.
Intersection-strict: The images of this set were the ones judged as relevant by both assessors.
Intersection-relaxed: The images of this set were the ones judged as relevant or partially relevant by both assessors.

This way, strict relevance and intersection sets can be considered as high-precision results, while relaxed relevance and union sets can be thought as results which promote higher recall.

4.1 Monolingual task

As stated before, the monolingual task consists on a set of queries in English, against a collection of image descriptions also in English.

In Figure 1 the recall vs. precision graph is presented for each of the five experiments submitted for this task. The values presented correspond to the evaluation of the results, comparing them with the Intersection-Strict relevance set (the more stringent one).

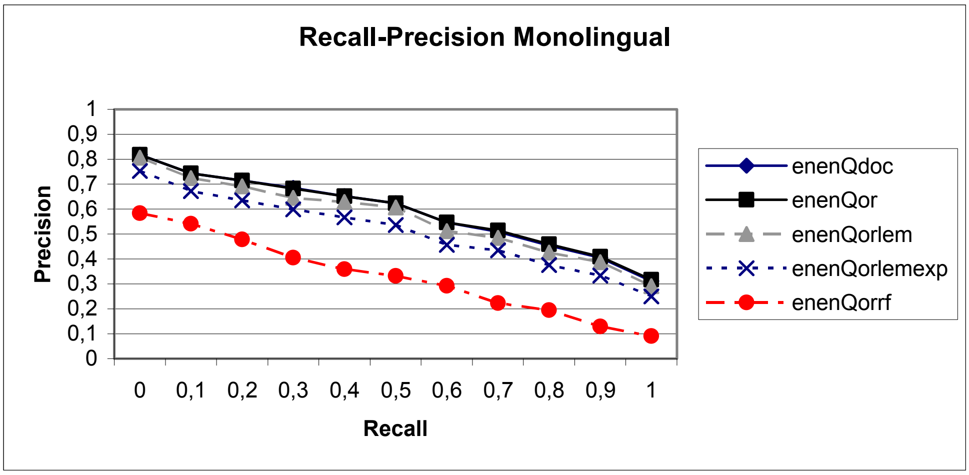

Figure 1. Recall-Precision graph for the Monolingual task

The first thing that can be noticed from this graph is the fact that the best runs have a quite high precision value, specially taking into account that image retrieval is a difficult task. In fact too high if compared with monolingual document retrieval results we have obtained in the monolingual tasks of CLEF2003. The explanation we find is that this year only four groups have taken part in ImageCLEF, and due to the way relevant sets were established (based basically on the submissions of each group), the actual cover of relevant documents was not as complete as should have been. That could be why so high values of precision have been obtained.

Other interesting aspect of the presented results is the fact that the run using blind relevance feedback leads to much worse results than all the other strategies. A possible explanation could be that the values used in the automatic relevance feedback were not appropriate to the kind of documents we were trying to retrieve. In fact, we used the top 250 terms of over the first 25 images retrieved. Given that each image has a mean length of the description field of 50 words, it becomes apparent that the number of relevant terms retrieved could be excessive. Therefore, instead of an aid to locate more relevant images, these terms only add noise that seriously penalize the overall performance.

It is worth mentioning that, instead of increasing the performance of the system, using any kind of term expansion (adding original query words or doing synonym expansion), only reduces the precision of the results. This could be due to the relatively low number of images of the collection, that doesn't make necessary to use word expansion to minimize the effect of heterogeneous descriptions that would arise in larger collections from different sources. Perhaps this strategy could be of interest in next ImageCLEF track, which, probably, will include larger collections.

Figure 2 represents the average precision of each submitted run among all topics, ordered from better to worse. This graph constitutes a simpler representation of the overall performance value for each experiment than the recall-precision graph, allowing to grasp in a single sight the quantitative differences of each approach. Again, the values presented are calculated considering the Intersection-Strict relevance set.

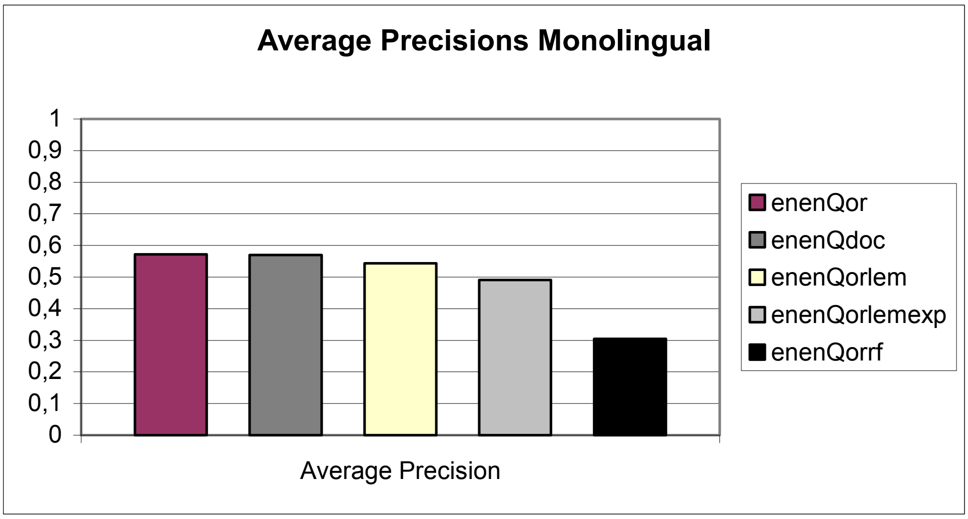

Figure 2. Precision comparison of different runs

As previously noticed in the recall–precision graph, it clearly states the poor performance of our relevance feedback experiment, and the similarity of the remainder experiments, specially the simple OR query approach (enenQor) and the query-indexing approach (enenQdoc).

Although only Intersection-Strict relevance sets have been mentioned in this section, differences with the other ones is subtle, apart from a slight increase of the overall precision in all cases due to the larger number of relevant documents they have.

4.2 Bilingual tasks

The bilingual tasks consist on the execution of queries in languages other than English, trying to retrieve relevant documents from a set of images described in English. Although queries in Spanish, Italian, German, French and Norwegian were available, we only took part in the first four languages.

Figure 1 shows the different precision vs. recall graphs obtained for each of the submitted runs and language pairs treated. In every case the values were obtained using the Intersection – Strict relevance set, being the strictest of all result sets provided.

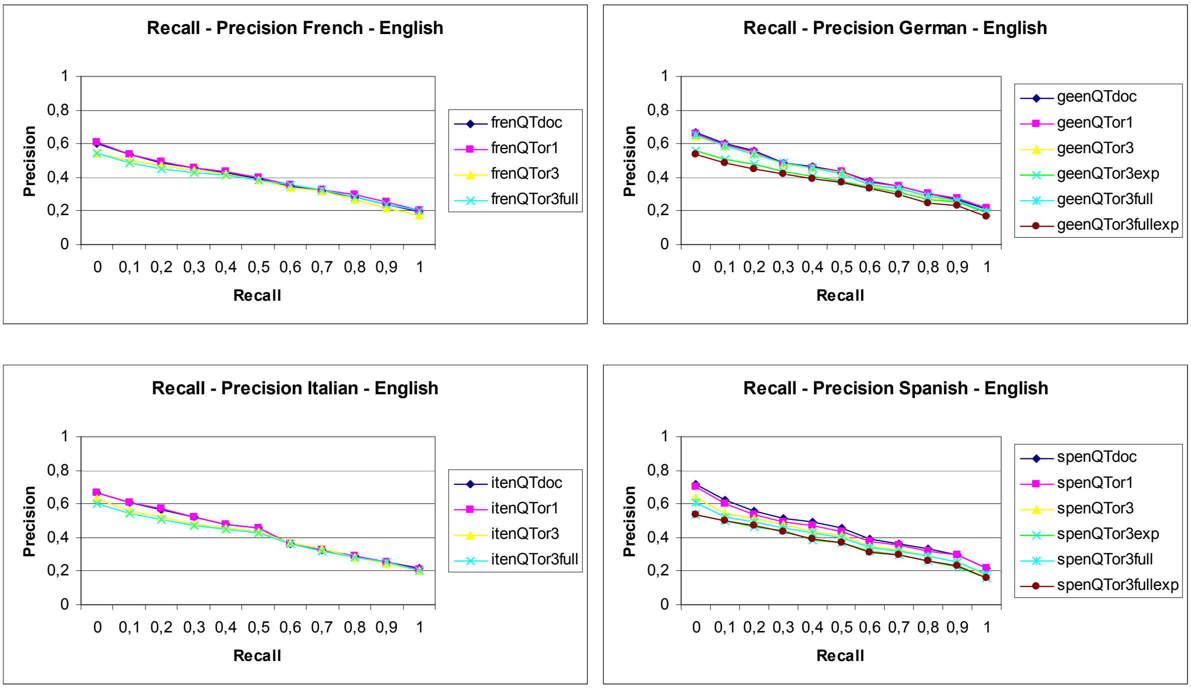

Figure 3. Recall - Precision graphs for bilingual tasks

Several conclusions can be extracted from these graphs. The more remarkable could be the similarity of QTdoc, QTor1, QTor3 and QTor3full experiments, being in all cases QTor1 and QTdoc the best. This is somehow consistent with the results obtained in the monolingual task, where the best performance was obtained by simple ORing the terms of the query (previously stemmed and removed stop words) –enenQor submission-, and by indexing the query as other image description and searching similar documents in the system –enenQdoc submision.

Another interesting aspect the graphs shows is that the use of more than one automatic translation has shown to be worse in our case than just using one of some quality (as the FreeTranslation has proved to be). It should be studied in more detail whether the use or ERGANE as the word by word translator was the cause of this lose of quality or was the simple fact of including word by word translation instead of only complete query ones. It is likely to be the second reason, since word by word translation always lead to wider queries that although increase recall, making precision worst. An example of this can be found in German to English and Spanish to English runs, in which synonym expansion in included (wider queries), leading, as expected, to worse precision values.

Another fact to point out is that precision values obtained in each task are quite similar, except for the French to English queries, which were slightly worse than the others. The explanation to this could be the worse French to English translations provided by FreeTranslation, or the use of different terms (hardest to translate) in the French queries.

Comparing the overall performance of the bilingual tasks with the monolingual one, a difference of about 10 to 15% arise, which is quite normal in typical CLIR nowadays. At least this is approximately the same value we have obtained in bilingual tasks of CLEF this year (as could be expected).

Figure 4 shows the average precision of each of the different runs submitted, considering the Intersection-Strict relevance set, as usual. The runs are ordered by descending precision and grouped by tasks.

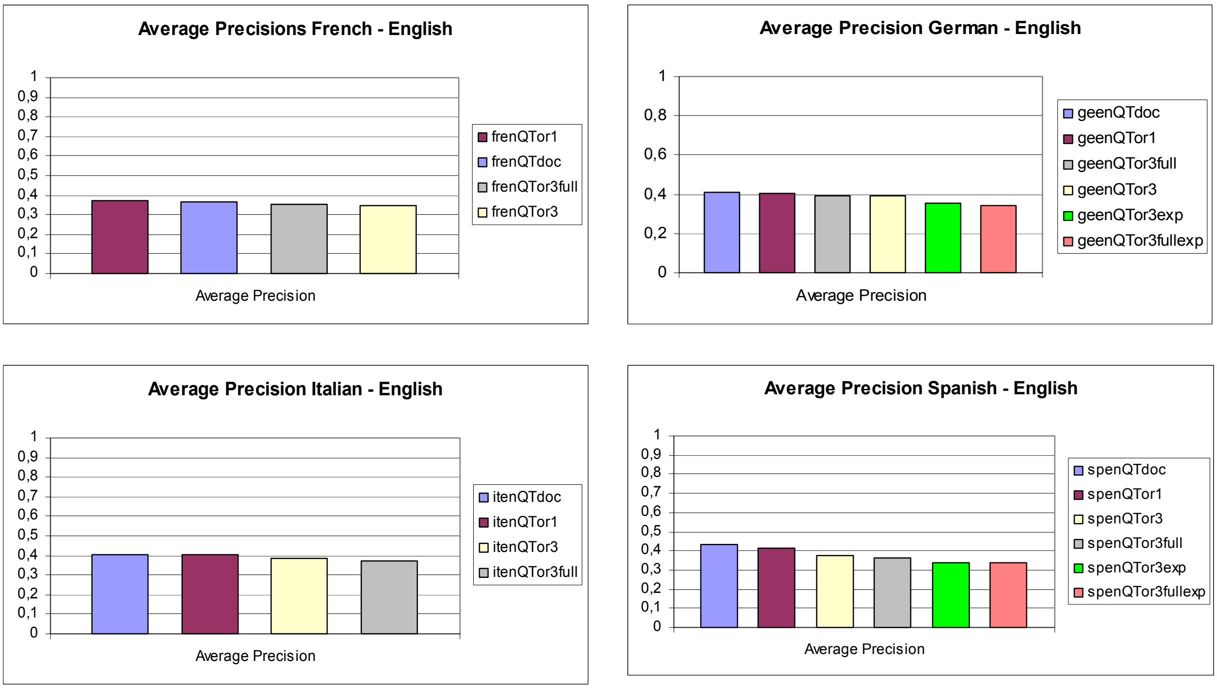

Figure 4. Precision comparison between runs

As in the case of the monolingual task, the results show little difference among different approaches, but consistently outperforming QTdoc and QTor1. It is once more apparent that our French to English retrieval has been slightly worse than the others, while the Spanish to English have obtained the best individual results (while not the best average results in all its runs).

5 Conclusions and future directions

The main conclusion that can be extracted taking into account the obtained results is that the simplest approaches studied (ORing terms and indexing the query and looking for similar documents) are the ones which lead to better results. Our main goal pursued with this first participation in the ImageCLEF track was to establish a starting point for future research work in the cross-language information retrieval applied to image (and in general other non-textual types of data that can be represented somehow by textual descriptions, such as video). Taking into account the obtained results, it stands out that there is much room for improvement both in monolingual and bilingual retrieval performance.

Also, despite the apparent bad results derived of performing synonym expansion, it looks like an interesting field to continue doing research on, especially for its likely application to wider and more heterogeneous collections.

References

[1] Baeza-Yates, R., Ribeiro-Prieto B., “*Modern Information Retrieval*”, Addison Wesley (1999).
[2] Karen Sparck Jones y Peter Willet, “*Readings in Information Retrieval*”, Morgan Kaufmann Publishers, Inc. San Francisco, California, 1997.
[3] “Free Translation”, www.freetranslation.com
[4] “Ergane Translation Dictionaries”, http://dictionaries.travlang.com
[5] “The Xapian Project”, www.sourceforge.net
[6] G.A. Miller. “***WordNet****: A lexical database for English*”. Communications of the ACM, 38(11):39—41, 1995.
[7]“The Porter Stemming Algorithm” page maintained by Martin Porter.
www.tartarus.org/~martin/PorterStemmer/