Exploiting Semantic Features for Image Retrieval at CLEF 2005

Martínez-Fernández, J.L.[2], Villena, J.[2,3], García-Serrano, Ana[1]
González-Tortosa, S.[1], Carbone, F.[1], Castagnone, M.[1]

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

joseluis.martinez@uc3m.es, jvillena@daedalus.es,
agarcia@isys.dia.fi.upm.es, sgonzalez@dia.fi.upm.es,
fcarbone@isys.dia.fi.upm.es, mcastagnone@isys.dia.fi.upm.es

Abstract

This paper presents the MIRACLE's team approach to text-based image retrieval at ImageCLEF 2005 adhoc task. The experiments defined this year try to use semantic information sources, like semantic dictionaries or text structure. For this purpose EuroWordnet has been considered and a new algorithm to extract synonyms from the semantic database has been developed. This new algorithm implementation is based on the proximity of words in the EuroWordnet tree and has been previously studied in [11]. On the other side, semantic information is implicitly included in the fields in which image descriptions are structured.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software. E.1 [Data Structures]; E.2 [Data Storage Representations]. H.2 [Database Management].

Keywords

Linguistic Engineering, Information Retrieval, text-based image retrieval, semantic data.

1 Introduction

ImageCLEF is the cross-language image retrieval track which was established in 2003 as part of the Cross Language Evaluation Forum (CLEF), a benchmarking event for multilingual information retrieval held annually since 2000. Images are language independent by nature, but often they are accompanied by texts semantically related to the image (e.g. textual captions or metadata). Images can then be retrieved using primitive features based on its contents (e.g. visual exemplar) or abstract features expressed through text or a combination of both.

Originally, ImageCLEF focused specifically on evaluating the retrieval of images described by text captions using queries written in a different language, therefore having to deal with monolingual and bilingual image retrieval (multilingual retrieval was not possible as the document collection is only in one language). Later, the scope of ImageCLEF widened and goals evolved to investigate the effectiveness of combining text and image for retrieval (text and content-based), collect and provide resources for benchmarking image retrieval systems and promote the exchange of ideas which will lead to improvements in the performance of retrieval systems in general.

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a company founded in 1998 as a spin-off of two of these groups. DAEDALUS is a leading company in linguistic technologies in Spain and is the coordinator of the MIRACLE team. This is the third participation in CLEF, after years 2003 and 2004 [5],[8],[12],[15],[18]. As well as bilingual, monolingual and cross lingual tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and GeoCLEF tracks.

This year a semantic driven approach to image retrieval has been tried. Semantic tools used have been: EuroWordnet [3] and textual image descriptions structure. A new implementation of a query semantic expansion has been developed, centered on the computation of closeness among the nodes of the EuroWordnet tree, where each node corresponds to a word appearing in the query. An expansion method based on the same idea was previously described in [11]. On the other hand, image captions have a predefined structure, each line of the text corresponds to a field. This information is exploited to build different indexes according to the type of field considered.

2 Semantic Expansion using EuroWordnet

EuroWordnet is a lexical database with semantic information in several languages. In the semantic level, for a given language, different relations have been defined among dictionary entries. These relations include: hyperonym, where links with more general concepts are defined, hyponym, where relations with more specific terms are included and synonym, where constructions grouping entries with the same meaning (named synsets) are built. All possible meanings for a given concept are part of the EuroWordnet data structure. So, as can be seen, a tree graph can be built using these semantic relations, and the distance among concepts in this tree can be used as a disambiguation method when expanding query expressions.

For example, the entry *bank* is defined in EuroWordnet as "*a financial institution that accepts deposits and channels the money into lending activities*" and also as "*sloping land (especially the slope beside a body of water)*" along with eight more different senses. The question arising is: how can be the word *bank* disambiguated when used as part of a query? The answer considered in this work is: by means of the rest of the words appearing with *bank* in the query. That is, some of the synonyms for the words appearing with the word *bank* will overlap with the synonyms of *bank*. If it does not happen hyponyms and hypernyms of the given words are considered, until some relations among the initial words are found. The senses which are not linked with the senses of other words appearing in the query expression can be discarded. Somehow, the main goal is to find one unique path, in the EuroWordnet tree, joining all the words that are present in the query.

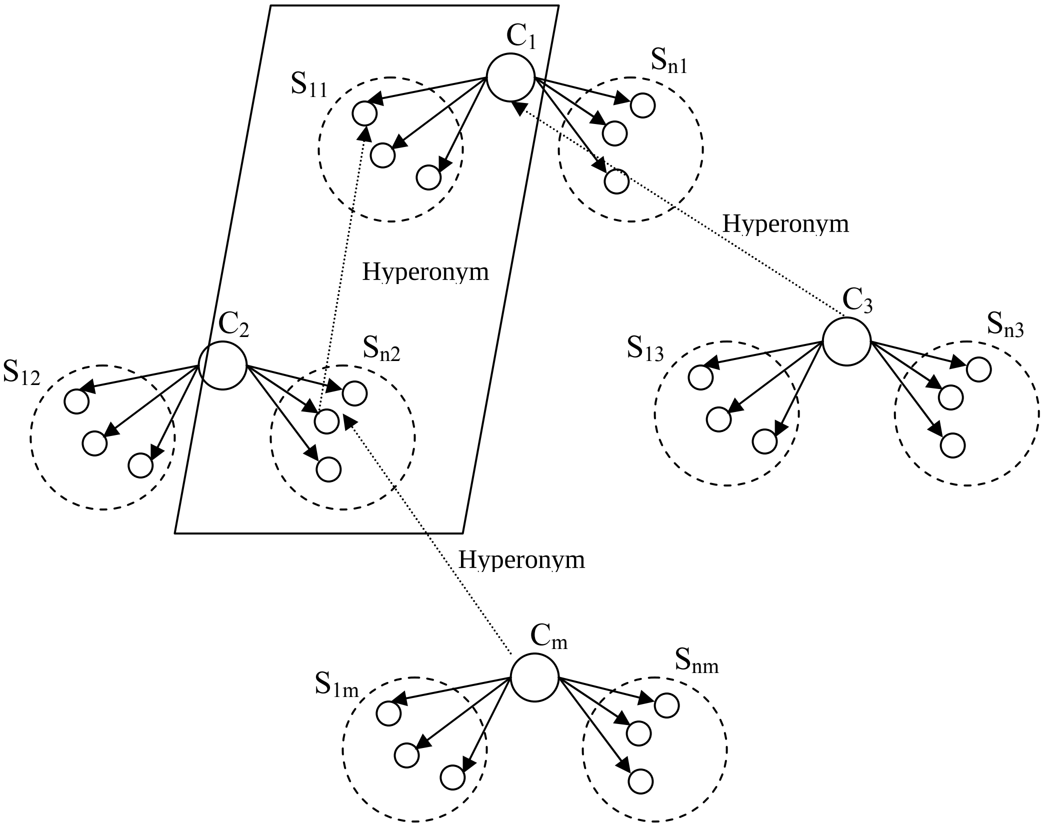

Figure 1. Hyperonym relations in EuroWordnet

The described situation is depicted in Figure 1. The dashed area corresponds to semantically related concepts, where the sense Sn2 for the concept C2 (appearing in the query) is related, by a hyperonym relation, with the sense S11 for the concept C1 appearing in the query. In this way, concepts C1 and C2 can be expanded including words in S11 and Sn2 sets, discarding the rest of senses, Sn1 and S12.

The described algorithm has been implemented using Ciao Prolog and an adaptation of the Dijkstra algorithm has been developed to compute the shortest way between two nodes. An efficient implementation of the expansion method has been pursued and, for this reason, not all possible paths among nodes are computed, a maximum of three jumps are allowed to limit execution times to an affordable value.

3 Morpho-Syntactic Processing

A more refined linguistic processing has been applied to the supplied captions. The availability of a tool to make morpho-syntactic analysis of English texts, based on the TreeBank tag set [14], allows for a deeper linguistic processing. This module is in charge of assigning a POS tag to each word, also identifying phrases appearing in a sentence. Negative particles present in sentences can then be identified, so specific and explicitly non-desired terms can be excluded when performing the search process. To identify these negative particles, an analysis of different sets of topics was carried out, obtaining a set of patterns to be matched against the input text. Terms obtained with this process were excluded from the documents to be retrieved by applying the corresponding operator provided by Xapian. This kind of processing could only be applied to English documents and several runs were submitted including this functionality, which has been always used together with the semantic expansion with EuroWordnet.

4 Exploiting image caption structure

The captions supplied for the St. Andrews image collection are divided in fields, each of them containing specific information such as short title, location, etc. Image textual descriptions are as shown in Figure 2. A total of 9 fields are defined for each caption, and only some of them are considered of interest for the defined retrieval tasks. Taking into account this structure, several indexes have been defined, one containing only image descriptions, another one with short title, one more with the photographer, another one with the places shown in the images, one with the dates when the pictures were taken and the last one with the proper nouns that have been identified in the image caption. In this way is possible to isolate pieces of information, allowing the retrieval of images based on specific data and mixing the results of different retrieval processes over distinct indexes if necessary. This year, again the Xapian search engine [19] has been used to index text representations for the image captions and the ability for this search engine to perform search processes combining independent indexes has been used.

Record ID	RMA-H.003430
Short Title	Sound of Jura, Garbh Reisa
Long Title	Seascape, from Craignish point towards Garbh Reisa (Reis).
Description	View over rock and grass hillocks to calm sea with string of islands; hills fading in far distance; cloudy sky.
Date	18 August 1933
Photographer	Robert Moyes Adam
Location	Argyllshire, Scotland
Notes	jf OS 52CONDN: Spots and scratches. PUBL: Leng/Thomson information on envelope, 1960. SCAN: Hairs.
Categories	[islands - coastal],[seascapes unclassified],[Argyll all views],[Collection - R M Adam]

Figure 2. Structure of image captions in the St. Andrews image collection

This information distribution allows for the assignment of semantic interpretation for each field and, with a minimum processing for the query, it is possible to search a specific entity over the right index. For example, several queries ask for images taken by a predefined photographer; a simple processing of the query allows for the identification of structures like "... taken by ..." where the name to be searched can be extracted and located over the picture author index. This strategy allows for a fine-grained search process that is supposed to provide better precision figures.

5 Experiments Description

This year, mono and bilingual experiments have been performed. In the monolingual experiments, a total of 17 executions have been submitted, while for the bilingual experiments 89 runs have been sent for 23 different source languages. As it is well known, an Information Retrieval process is divided in two main subtasks, indexing and searching. To obtain the best retrieval performance both subtasks can be parameterized, taking always into account that index terms and search terms must be represented using a common model (i.e.: in some situations, the same alterations introduced for index terms must be produced for search terms). Regarding the query processing subtask, the following features that can be included:

Source Field: The topics used in the ImageCLEF track are divided in two main fields, a title (a short description of the aim of the topic) and a narrative (with a more detailed description of the topic purpose). Depending on which of these fields are used, several experiments can be defined. In our case, experiments where only the title field for the query has been used are marked with 't0'; if only the narrative field has been considered, the name for the run is marked with 'd0' and, when both fields are used together, a 'td' is included in the name of the run.

- *Baseline*: This is the basic and simplest approach whose results constitute the record to break. The transformations performed with the input text are: a basic parser is in charge of dividing the text in words, then all words are normalized (by lowercasing every letter and removing special characters), stopwords are removed and, finally, the stem for each word in the query is obtained. The resulting words are then used to search the indexes.
- *Query Expansion*: As explained in section 2, a semantic expansion algorithm has been implemented this year, to include in the query semantically related concepts for the provided words. This method is applied at the output of the baseline process (but passing the stemming step to the end, i.e., to the output of the query expansion module).
- *Linguistic Processing*: This component (described in section 3) is in charge of obtaining the linguistic analysis of the text contained in the query. As already mentioned this analysis is in charge of selecting only the nouns appearing in the topic and of identifying which words should be excluded from the query. This module is always used in combination with the Query Expansion component.
- *Combination Operator*: There are two possible ways of joining together the words of the query with the expanded terms. The simplest one is using the OR operator to combine every pair of words. One more complex way of joining terms is considering the OR operator to join synonyms for a word and the AND operator to join sets of synonyms for different words.
- *Proper noun module*: A simple proper noun detection module, based on a finite state automaton, has also been applied to image captions and topics. Some attempts to work only with proper nouns identified in the query and in image descriptions. Because of the poor recall and precision figures obtained with former ImageCLEF data sets, none of these runs were sent.

Until this point, the different transformations applied to the query processing task have been described. Now, the processes followed in the indexing subtask are explained. Several indexes have been built with the image captions collection, where different data arrangements were defined. When more than one of these indexes was targeted as part of the same search process, the ability of the Xapian search engine to perform queries over several databases[1] at the same time was exploited. A total of seven indexes were built:

- *Caption index*: All information contained in the image caption was used to build a unique index.
- *Title index*: The titles included in image textual descriptions were indexed in the same database.
- *Description index*: Only the description field of the image caption was used to build an index.
- *Author index*: All contents of *Photographer* fields present in image captions were indexed as part of the same database.
- *Place index*: Words appearing in the *Location* field of image captions were used to build an index.
- *Date index*: Words and dates present in the *Date* field of the captions were indexed as part of the same database.
- *Proper Nouns index*: All proper nouns detected in any of the fields defined for the image captions are included in this index. Results presented in section 6 do not include experiments involving this index because it was rejected due to the low precision and recall values obtained.

When fields with plain text content are treated (such as the 'Description' or ''Title' attributes) the same parser, stopwords removing, accented characters substitution and stemming processes than the ones applied to the queries are considered.

[1] Note that the term 'database' is being used with the same sense that 'index'

Taking into account these descriptions, the nomenclature followed for the submitted experiments is depicted in Figure 3. The possible values for each field are:

- the 'Query field used' can take values: 't0', when only the query title is used, 'd0', when only the narrative field is used, and 'dt' when both title and narrative are used to build the search expression for the search engine.
- the 'Linguistic processing applied to the query' can have values: 'base', when the processes for the baseline are applied (i.e.: parsing, stopword filtering, special characters substitution and lowercasing and stemming); 's', when the module to obtain the morphosyntactyc analysis for the query is used; 'e', when the semantic expansion based on EuroWordnet is applied; 'o', when the operator to combine the expanded words is OR; 'a' when the operator to join expanded query words is a combination of OR operators with AND operators; 'pn', when proper nouns are identified in the text.
- the 'Index used' field identifies which index (or indexes) is (are) used to retrieve images. The possible values are: 't0', if only the titles of the captions are indexed, 'd0', when only the descriptions for the captions are searched, 'dt', when both titles and descriptions constitute a unique index, 'attr', if indexes for the different captions fileds are used (the identified fields are: text, author, date, place), and finally 'allf', when a unique index with the content of all fields is used.
- the 'Source Language' part identifies the language in which the query is supplied. In monolingual experiments it is English, but for bilingual experiments it can it can identify one from 22 different languages (Bulgarian, Croatian, Czech, Dutch, English, Finnish, Filipino, French, German, Greek, Hungarian, Italian, Japanese, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish - Latinamerica, Spanish - Spain, Swedish, Turkish and Simplified Chinese.
- the 'last part, denoted Target Language', identifies the language in which the image captions collection is written. Until now, the target language is always English.

For bilingual experiments, were the source language is other than English, different translation tools have been used to transform the original query texts to English. Among these translation tools is with mentioning Systran 5.0 [15], FreeTranslation [4] and TranExp InterTran [16]. Once the queries have been translated, the following processes are the same than the ones used in the monolingual experiments.

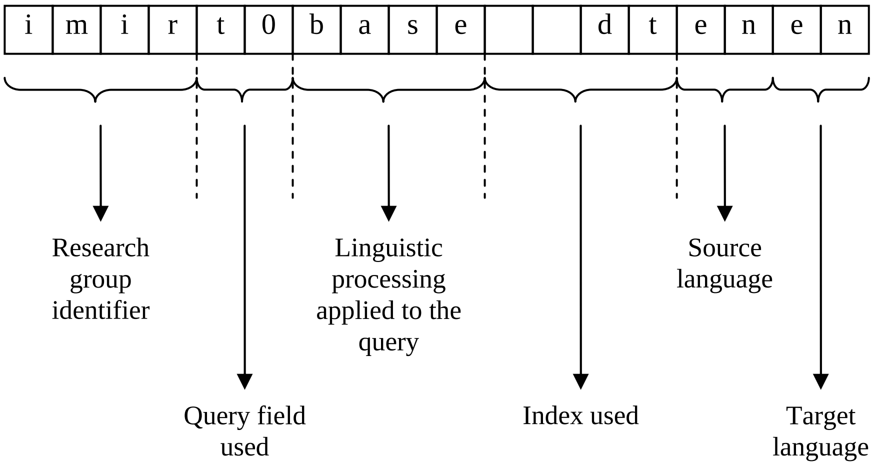
i m i r t 0 b a s e d t e n e n
Research group identifier
Linguistic processing applied to the query
Source language
Query field used
Index used
Target language

Figure 3. Experiments nomenclature

6 Results

Obtained results can be divided according to the languages involved in the retrieval process.

¡Error! No se encuentra el origen de la referencia. shows the Medium Average Precision (MAP) for the monolingual experiments presented this year by the Miracle group. The best monolingual result is obtained for experiment 'imirt0attren', where the title for the topic is processed with the baseline procedure (parsing, normalizing words, stopwords removal and stemming) and the built query is performed against the combination of attribute indexes (text, place, author, date). The MAP for this experiment is 37%, not far from the next one 'imirt0allfen'. It is worth mentioning that these figures are not conclusive, a programming error in the combination of the different indexes introduced duplicate entries in the final result list. These duplicate results were simply deleted from the final result list and lowering precision and recall rates. To the time of writing, it has not be possible to repeat the experiments to produce new runs without duplicates.

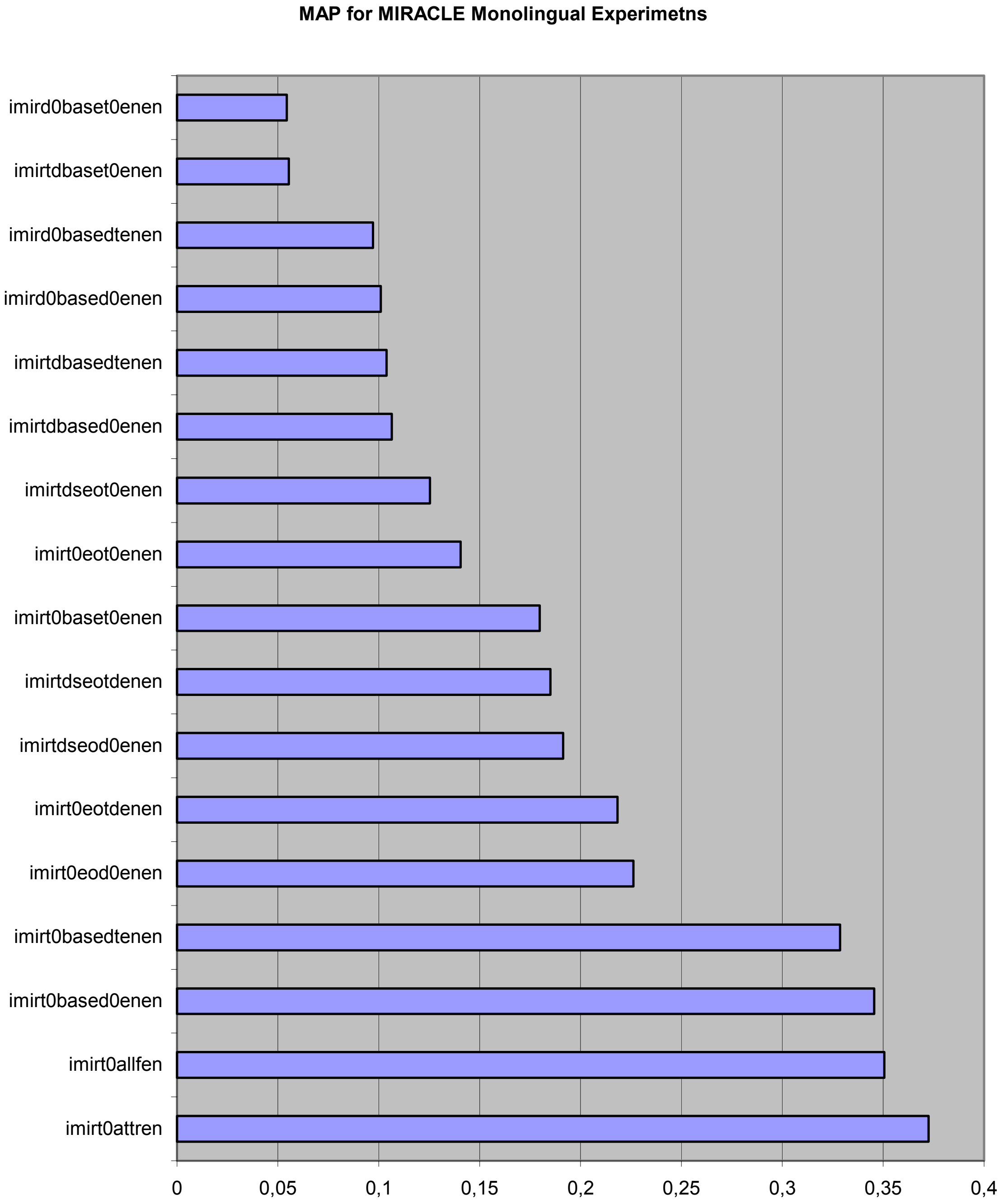

Figure 4. Medium Average Precision for MIRACLE's monolingual runs at Image CLEF 2005

Results for bilingual experiments are also very interesting. In **¡Error! No se encuentra el origen de la referencia.**, a graph showing the differences among the experiments for each language is depicted. The MAP precision values for the best result for each language are compared. The best bilingual MAP result is 31%, and it is reached for the Portuguese language. Comparing with the best monolingual result, a difference of around 7% in MAP value can be seen.

As already tested in previous campaigns, the translation process between languages introduces a lot of noise, decreasing the precision of the retrieval process. The process followed in the 'imirt0attrpt' experiment is equivalent to the one applied in the best monolingual run, but including a previous translation step using the previously mentioned translators. That is, the topic title is translated from Portuguese into English and then parsed, normalized, stopwords are removed and the rest of words are stemmed. The words forming the query are ORed and searched against the combination of attribute indexes (text, place, author, date). Of course, the previously explained problem with duplicate results in the final list also applies to the bilingual runs submitted.

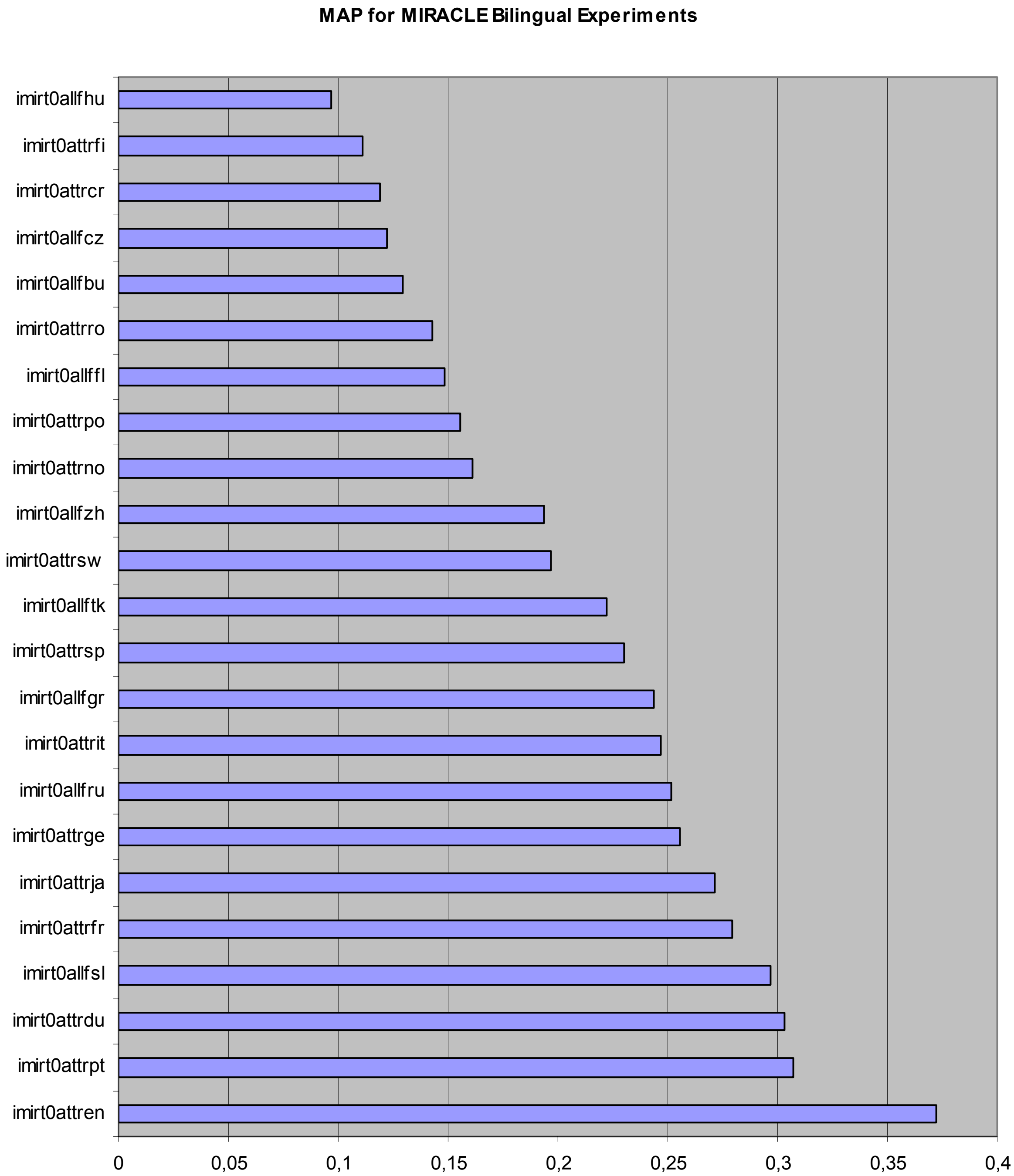

Figure 5. Medium Average Precision for MIRACLE's bilingual runs at Image CLEF 2005

It can also be observed that the MIRACLE team has been the only participant for some target languages such as Bulgarian, Croatian, Czech, Filipino, Finnish, Hungarian, Norwegian, Polish, Romanian and Turkish.

7 Conclusions and Future Works

The results shown in the previous section can lead us to some preliminary conclusions. The best results are obtained for the approach where information in captions fields is isolated. The query expansion based on EuroWordnet dos not lead to better results, although more accurate synonyms are selected by applying the new expansion algorithm which tries to disambiguate different senses for the words appearing in the topic text. The reason can be related with the number of words constituting the query once the expansion is performed; too much information to perform a quality search. This idea is reinforced if we take a look to the MAP produced by the experiment 'imirtdbasedtenen', where the same process is followed than in 'imirt0basedtenen' but also using the narrative field of the topic. In this situation a 10% MAP is obtained, a 20% worst than the experiment where only the title of the topic is used.

On the other hand, no definitive conclusions can be drawn until the experiments are repeated without producing duplicate results in the final list.

Future works in text-based image retrieval could be devoted to the use of the trie-based indexing tool available among the utilities developed by the MIRACLE team and, on the other hand, to the improvement of the semantic expansion algorithm which, although it has not been proved to be useful for this task, seem to be accurate if the obtained expansions are visually checked.

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 and also by the European Union with the funding of NEDINE project in the e-Content programme.

Special mention to our colleagues of the MIRACLE team should be done (in alphabetical order): José Carlos González-Cristóbal, Ana González Ledesma, José Miguel Goñi-Menoyo, José Mª Guirao, Sara Lana-Serrano, Paloma Martínez-Fernández, Ángel Martínez-González, Antonio Moreno Sandoval and César de Pablo Sánchez.

References

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

[2] Aoe, Jun-Ichi; Morimoto, Katsushi; Sato, Takashi. An Efficient Implementation of Trie Structures. Software Practice and Experience 22(9): 695-721, 1992.

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

[4] Free2Translation. Free text translator. On line http://www.freetranslation.com [Visited 20/07/2005].

[5] Goñi-Menoyo, José M; González, José C.; Martínez-Fernández, José L.; and Villena, J. MIRACLE’s Hybrid Approach to Bilingual and Monolingual Information Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005 (to appear).

[6] Goñi-Menoyo, José M.; González, José C.; Martínez-Fernández, José L.; Villena-Román, Julio; García-Serrano, Ana; Martínez-Fernández, Paloma; de Pablo-Sánchez, César; and Alonso-Sánchez, Javier. MIRACLE’s hybrid approach to bilingual and monolingual Information Retrieval. Working Notes for the CLEF 2004 Workshop (Carol Peters and Francesca Borri, Eds.), pp. 141-150. Bath, United Kingdom, 2004.

[7] Goñi-Menoyo, José Miguel; González-Cristóbal, José Carlos and Fombella-Mourelle, Jorge. An optimised trie index for natural language processing lexicons. MIRACLE Technical Report. Universidad Politécnica de Madrid, 2004.

[8] Martínez-Fernández, José L.; García-Serrano, Ana; Villena, J. and Méndez-Sáez, V.; MIRACLE approach to ImageCLEF 2004: merging textual and content-based Image Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

[9] Martínez, José L.; Villena, Julio; Fombella, Jorge; G. Serrano, Ana; Martínez, Paloma; Goñi, José M.; and González, José C. MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 210-219. Springer, 2004.

[10] Martínez, J.L.; Villena-Román, J.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.): Evaluation of MIRACLE approach results for CLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[11] Montoyo, A., *"Método basado en marcas de especificidad para WSD"*, In Proceedings of SEPLN, nº 24, September 2000.

[12] de Pablo, C.; Martínez-Fernández, J. L.; Martínez, P. and Villena, J. *miraQA*: Initial experiments in Question Answering. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

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

[14] B. Santorini, "*Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision*)," Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Tech. Rep. MS-CIS90 -47, Line Lab 178, 1990, ftp://ftp.cis.upenn.edu/pub/treebank/doc/ manual/root.ps.gz.

[15] SYSTRAN Software Inc., USA. SYSTRAN 5.0 translation resources. On line http://www.systransoft.com [Visited 13/07/2005].

[16] Translation Experts Ltd. InterTrans translation resources. On line http://www.tranexp.com [Visited 28/07/2005].

[17] Villena, Julio; Martínez, José L.; Fombella, Jorge; G. Serrano, Ana; Ruiz, Alberto; Martínez, Paloma; Goñi, José M.; and González, José C. Image Retrieval: The MIRACLE Approach. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 621-630. Springer, 2004.

[18] Villena-Román, J.; Martínez, J.L.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.); MIRACLE results for ImageCLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

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