MIRACLE team report for ImageCLEF IR in CLEF 2006

Martínez-Fernández, José Luis[1], Villena, Julio[1,3], García-Serrano, Ana[2], Martínez, Paloma[1]

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

joseluis.martinez@uc3m.es, jvillena@daedalus.es,
agarcia@isys.dia.fi.upm.es, paloma.martinez@uc3m.es

Abstract

The hypothesis which this paper tries to validate is that text based image retrieval could be improved by the use of semantic information, by means of an expansion algorithm and a module specifically designed to exclude common words and negated words from queries. The expansion algorithm applies specification marks to disambiguate words making use of WordNet [13]. An implementation of this algorithm has been developed for these experiments. On the other hand, the module in charge of removing common words and detecting negated words has also been specifically developed for this work. However, after an initial evaluation, none of these modules led to an improvement in the retrieval quality compared to the baseline experiment, which consists on the indexing of nouns present in image captions, without no further preprocessing.

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, Image Retrieval, Semantic Expansion, WordNet, Word Sense Disambiguation

1 Introduction

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a leading company in linguistic technologies in Spain, spin-off of two of these groups, and the coordinator of the MIRACLE team. This is our fourth participation in CLEF, after years 2003, 2004 and 2005. As well as bilingual, monolingual and robust multilingual tasks [3], the team has participated in the ImageCLEF [9], Q&A [2], WiQA, iCLEF [16], WebCLEF [12] and GeoCLEF [6] tracks.

Following the structure of previous campaigns, ImageCLEF task has been divided again in several subtasks as described in [4]. This year, the MIRACLE team has only taken part in the ImageCLEFPhoto task. The main goal this year for our team was to make use of the new image collection, IAPR, described in [4], and, using the new textual captions provided, try to make a deep linguistic analysis of these captions in order to build some kind of semantic representation of the text. This representation uses Charniak's parser and is also based on WordNet. The main focus has been put on the implementation of the disambiguation algorithm proposed by [13], with some consideration to make it more easy to develop. In addition, a basic query analyzer has been added to produce a linguistic analysis of topics and filter some common expressions and words.

Topics proposed by the organization are divided in three sections, a title, a narrative, i.e. a longer description of the topic, and a set of images that could be used by a Content Based Image Retrieval (CBIR) system to perform the search. Image captions in the IAPR collection have different fields and, in the MIRACLE approach, only title and description fields have been extracted and indexed separately. Taking into account available elements, several experiments have been executed, which are described in section 4.

Regarding the multilingual dimension of the proposed task, although the IAPR provides captions in English and German, only the English language has been considered in our experiments. Topics have also been provided in several languages and the MIRACLE team has submitted runs for Japanese, Simplified Chinese, Russian, Polish and English.

2 Topic analysis and semantic expansion as implemented for CLEF 2006

Our previous experiences in image retrieval processes [9][10] have shown that, in practical, current indexing techniques have reached their precision and recall upper limit. To surpass this point, we think that semantic information should be considered in the retrieval process. For this reason the system developed this year includes an implementation of a WordNet based semantic expansion method that uses specification marks [13] (adapted to version 2.1 of WordNet) and a topic analysis module, which is intended to detect common words and to filter out words introduced by negation expressions.

The semantic expansion method was defined to disambiguate words appearing in WordNet, using the context of the word to select the correct sense among the set of senses assigned by WordNet. Intuitively, the idea is to select the sense pertaining to the hyperonym of the word that includes the greater number of senses for the words appearing in the context. The main objective is to include only the synonyms corresponding to the correct sense or the word instead of adding all synonyms for all senses of a word. Not every word can be disambiguated by applying this algorithm. Thus, to increase the number of correctly disambiguated words, three heuristics were identified. Only one of this heuristics has been included in our implementation, the *Definition Heuristic,* which discards senses whose gloss does not include any word present in the context. In CLEF 2005 we tried to use a variation of this algorithm, which was intended to disambiguate word pairs, but no optimal results were obtained and, so, this year we decided to use our implementation of the original algorithm. An example of the result of the expansion process applied to the title of topic 30 is shown in Figure 1.

WordNet based Semantic Expansion method applied to the title of Topic 30: "room with more than two beds"		
Filtered nouns	**Semantic Expansion without WSD**	**Semantic Expansion with WSD**
beds, room	beds, furniture, "piece of furniture", "article of furniture", plot, "plot of ground", patch, bottom, "natural depression", depression, stratum, layer, sheet, "flat solid", surface, foundation, base, fundament, foot, groundwork, substructure, understructure, room, area, way, "elbow room", position, "spatial relation", opportunity, chance, gathering, assemblage	beds, furniture, "piece of furniture", "article of furniture", room, area

Figure 1. Example of the semantic expansion method.

On the other hand, a topic analysis module has been developed to make a deeper selection of terms to search against the index. This module is used to filter out common words and expressions as well as to detect negation structures. Thus, phrases like *"Relevant images will show"* are excluded and words accompanying phrases like *"are not relevant"* are expressed with a negation symbol "-" which is interpreted by the search engine. When negations are interpreted, not every word taking part in the sentence is excluded, only those words that do not appear in an affirmative form are negated. An example of the result is shown in Figure 2.

Original narrative field (topic 50)	**Search engine query after topic analysis**
Relevant images will show an interior view of a church or cathedral. Images showing exterior views of churches or cathedrals are not relevant. Interior views of other buildings than fanes are not relevant.	church, cathedral, view, photos, -images, -exterior, -buildings,-fanes

Figure 2. Example of the topic analysis module.

These are the two main components included in the MIRACLE image retrieval system tested in this campaign. The following section describes system architecture. For both subsystems, a linguistic analysis of the topic text is needed. For this purpose, the Charniak's parser [1] has been integrated. To tokenize the text, i.e. to divide the text in sentences, the LingPipe tools [7] have been used.

3 System Architecture

The software architecture for the text based image retrieval system is depicted in Figure 3. As can be seen, the approach is based on the integration of modules that can be optionally activated in order to configure the different experiments to be submitted. The retrieval process is divided in two tasks: the indexing process, in charge of building the indexes to be searched. As already mentioned the used search engine is Xapian [18] and the indexing options have been the usual ones applied by this system, tokenization and stemming of words to be indexed. The *Text Extractor* component depicted in Figure 3 takes the XML image captions and extracts the content of each field. Then, the extracted text is stored in the proper index. During this process, a basic transliteration of characters is made, to avoid problems with special characters.

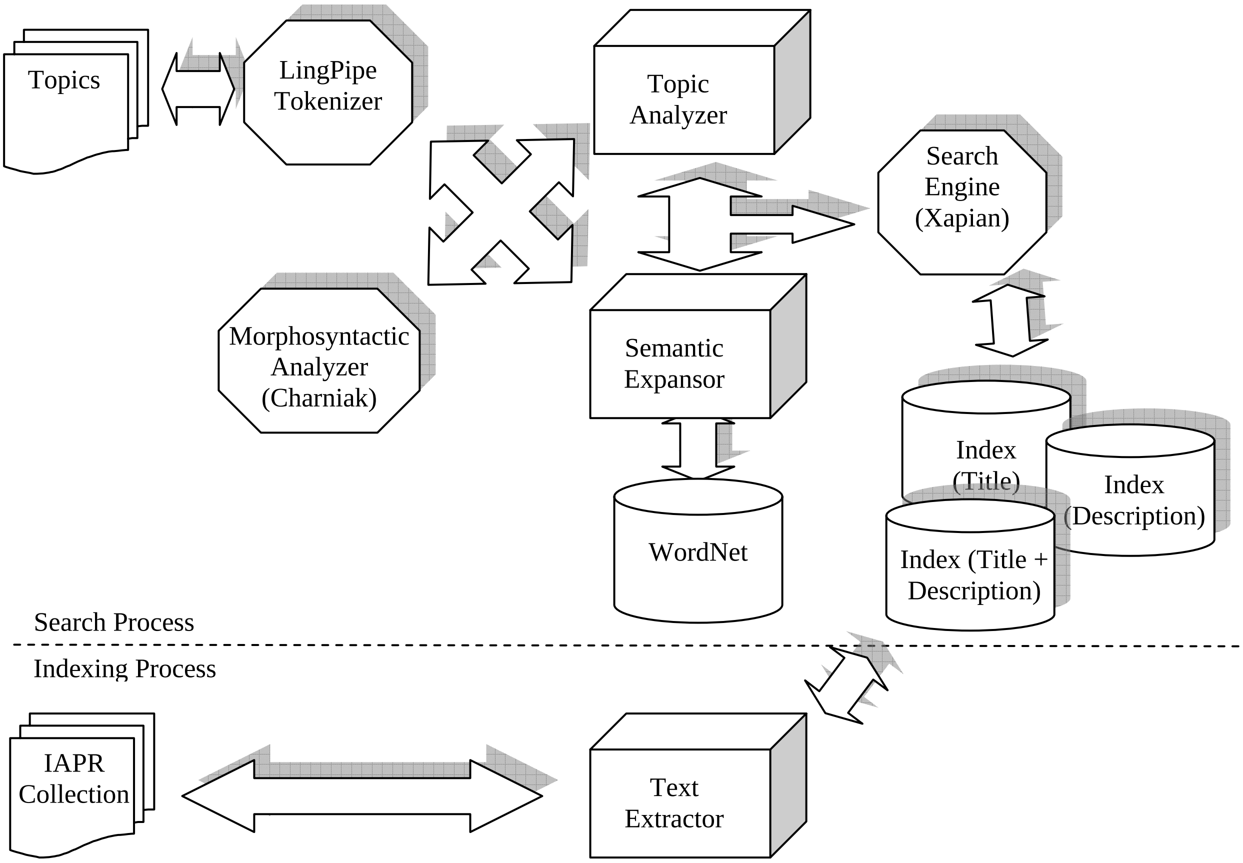

Figure 3. Architecture for the CLEF 2006 text based image retrieval system

Three different indexes are built: one containing the titles of the image captions, another one containing only the descriptions present in the image captions and the last one mixing titles and descriptions. The search process is devoted to the construction of the query to be executed on the previously built indexes. During this process several tasks are performed:

- *Tokenization*. The LingPipe software [7] is used to divide the text in the basic operation unit, which is considered to be the sentence. The identification of sentences in the input topic text is always performed.
- *Morphosyntactic Analysis*. The Charniak's parser [1] is used to obtain the morphosyntactic analysis of each sentence. These analyses are stored in a database to make them easier to manage and they constitute the input for the following processes.
- *Topic Analysis*. Topic analysis is applied to every input topic. Two options can be selected: *noun*, to extract only words tagged as nouns during the parsing process, or *common*, to filter out common expressions like *"Relevant images will show"* and to exclude words introduced by negation structures like *"[...] are not relevant"*. Only words tagged as nouns are considered and the "-" operator available in Xapian to exclude words is used. It is also possible to mark the topic section to be used: title, narrative or both.
- *Semantic Expansion*. Optionally, semantic expansion can be applied to the output of the Topic Analysis module. This semantic expansion is based on WordNet and the previously described disambiguation method is applied to select the synonyms to be included in the query. Words to be excluded from the query are ignored by the expansion algorithm and the disambiguation process is performed taking into account the scope defined by one sentence.

Different combinations of these modules configure the experiments that have been performed and submitted.

4 Defined experiments

Different modules have been selected to define experiments. Although the IAPR collection is available in two languages, English and German, only the English target language has been considered. On the other hand, four different query languages have been tested: English, Japanese, Polish, Russian and Traditional Chinese. Names and descriptions of runs are included in Table 1. The column *"Topic Part"* marks the field of the topic that have been processed: if the label *"Title"* is included, it means that only the title fragment of the topic has been used, whereas if the label *"Title+Narrative"* is given, it means that both fields of the topic have been processed. Bilingual runs are always marked with the label *"Title"* because this is the only field provided in topic in languages distinct than English. Values for the *"Topic Analysis"* column can be *"noun"* or *"common"* as described in the previous section. The *"Expansion"* columns takes *"Yes"* value if the Semantic Expansion module has been used or *"No"* in other case. Finally, the *"Index"* column points out which section of the image caption is used, the title fragment, the description fragment or both.

Table 1: Text-based experiments

Run Name	Topic Language	Topic Part	Topic Analysis	Expansion	Index
miratnntdenen	English	Title+Narrative	Noun	No	Title+Desc.
miranntdenen	English	Narrative	Noun	No	Title+Desc.
miranctdenen	English	Narrative	Common	No	Title+Desc.
miratnctdenen	English	Title+Narrative	Common	No	Title+Desc.
miratncdenen	English	Title+Narrative	Common	No	Desc.
miratnndenen	English	Title+Narrative	Noun	No	Desc.
miranndenen	English	Narrative	Noun	No	Desc.
mirancdenen	English	Narrative	Common	No	Desc.
miratctdjaen	Japanese	Title	Common	No	Title+Desc.
miratntdjaen	Japanese	Title	Noun	No	Title+Desc.
miratctdplen	Polish	Title	Common	No	Title+Desc.
miratctdzhsen	Trad. Chinese	Title	Common	No	Title+Desc.
miratntdzhsen	Trad. Chinese	Title	Noun	No	Title+Desc.
miratntdplen	Polish	Title	Noun	No	Title+Desc.
miratctdruen	Russian	Title	Common	No	Title+Desc.
miratnndtdenen	English	Title+Narrative	Noun	Yes	Title+Desc.
mirannd tdenen	English	Narrative	Noun	Yes	Title+Desc.
miranndddenen	English	Narrative	Noun	Yes	Desc.
miratnnddenen	English	Title+Narrative	Noun	Yes	Desc.
miratndtdjaen	Japanese	Title	Noun	Yes	Title+Desc.
miratndtdplen	Polish	Title	Noun	Yes	Title+Desc.
miratncdtdenen	English	Title+Narrative	Common	Yes	Title+Desc.
miratncddenen	English	Title+Narrative	Common	Yes	Desc.
mirancddenen	English	Narrative	Common	Yes	Desc.
mirancdtdenen	English	Narrative	Common	Yes	Title+Desc.
miratndtenen	English	Title+Narrative	Noun	Yes	Title
miratndtdruen	Russian	Title	Noun	Yes	Title+Desc.
miratndtdzhsen	Trad. Chinese	Title	Noun	Yes	Title+Desc.

Besides, two runs combining textual and content based indexing have been submitted. For this purpose, the content image indexing facilities of Lucene [8] have allowed the construction of an index with image contents. The three images supplied with the topics have been searched over the image index and the partial results list combined by adding obtained normalized relevances. The final result list has been combined again with the textual result list. Table 2 shows the description of these two runs:

Table 2: Mixed visual and textual experiments

Run Name	Topic Language	Topic Part	Topic Analysis	Expansion	Index
miratnntdienen	English	Title+Narrative	Noun	No	Title+Desc.
miratncdtdienen	English	Title+Narrative	Common	Yes	Title+Desc.

Obviously, there are more possible combinations, but these ones where considered the most appropriate according to previous experiences. For example, using only image titles of captions is not useful because a poor description of each caption can be obtained by the indexer. Besides, if only topic titles are used as the input for the search process, the characterization of the query built by the search engine is weak and no good results are usually obtained. Next section provides precision and recall measures for these experiments and a qualitative explanation for them.

5 Experiment results

Obtained results are somehow discouraging. Table 3 shows the MAP measure for the submitted runs.

Table 3: MAP figures for text-based experiments

Run Name	Topic Language	Topic Part	Topic Analysis	Expansion	Index	MAP
miratnntdenen	English	Title+Narrative	Noun	No	Title+Desc.	0,2009
miranntdenen	English	Narrative	Noun	No	Title+Desc.	0,1960
miranctdenen	English	Narrative	Common	No	Title+Desc.	0,1875
miratnctdenen	English	Title+Narrative	Common	No	Title+Desc.	0,1866
miratncdenen	English	Title+Narrative	Common	No	Desc.	0,1375
miratnndenen	English	Title+Narrative	Noun	No	Desc.	0,1361
miranndenen	English	Narrative	Noun	No	Desc.	0,1352
mirancdenen	English	Narrative	Common	No	Desc.	0,1314
miratctdjaen	Japanese	Title	Common	No	Title+Desc.	0,1252
miratntdjaen	Japanese	Title	Noun	No	Title+Desc.	0,1252
miratctdplen	Polish	Title	Common	No	Title+Desc.	0,1075
miratctdzhsen	Trad. Chinese	Title	Common	No	Title+Desc.	0,1041
miratntdzhsen	Trad. Chinese	Title	Noun	No	Title+Desc.	0,1041
miratntdplen	Polish	Title	Noun	No	Title+Desc.	0,1010
miratctdruen	Russian	Title	Common	No	Title+Desc.	0,0909
miratnndtdenen	English	Title+Narrative	Noun	Yes	Title+Desc.	0,0172
mirandtdenen	English	Narrative	Noun	Yes	Title+Desc.	0,0157
mirannddenen	English	Narrative	Noun	Yes	Desc.	0,0157
miratnnddenen	English	Title+Narrative	Noun	Yes	Desc.	0,0155
miratndtdjaen	Japanese	Title	Noun	Yes	Title+Desc.	0,0103
miratndtdplen	Polish	Title	Noun	Yes	Title+Desc.	0,0091
miratncdtdenen	English	Title+Narrative	Common	Yes	Title+Desc.	0,0084
miratncddenen	English	Title+Narrative	Common	Yes	Desc.	0,0082
mirancddenen	English	Narrative	Common	Yes	Desc.	0,0077
mirancdtdenen	English	Narrative	Common	Yes	Title+Desc.	0,0072
miratndtenen	English	Title+Narrative	Noun	Yes	Title	0,0069
miratndtdruen	Russian	Title	Noun	Yes	Title+Desc.	0,0038
miratndtdzhsen	Trad. Chinese	Title	Noun	Yes	Title+Desc.	0,0034

As can be concluded from these figures, the expansion modules does not produce any improvement, on the contrary, a decrease of 18 % in MAP is observed. Although the application of expansion methods have not been definitely proved to increase precision figures, the great decrease produced in these experiments is likely due to a bug in the implementation. The code and partial evaluations of the expansion algorithm are going to be reviewed to determine if it is working in the proper way. On the other hand, when the topic analysis module is used to analyse negation expressions, a decrease in MAP is measured. This is not a strange result, taking into account the complexity of topics. When the topic and image caption sections used in the retrieval process are regarded, one can conclude that if greater amounts of text are used in both topic and caption better precision is obtained. Finally, as observed in previous bilingual retrieval experiments, when the language of topics is different of the language of the document collection an average 10% decrease in MAP is produced. This is due to the noise introduced by the translation step needed in these situations.

Table 4 shows MAP values for the experiments where content based image retrieval is used to support textual retrieval. As concluded in past experiments, content based partial results have no effect in the retrieval precision.

Table 4: MAP figures for mixed visual and textual retrieval experiments

Run Name	Topic Language	Topic Part	Topic Analysis	Expansion	Index	MAP
miratnntdienen	English	Title+Narrative	Noun	No	Title+Desc.	0,2016
miratncdtdienen	English	Title+Narrative	Common	Yes	Title+Desc.	0,0084

6 Conclusions

One direct conclusion from the previous section is that the experiment considered as the baseline could not be improved. Although a deeper exploration of results and processes have to be carried out, initially seems to be due to a improper operation of the expansion module. Besides, it is worth mentioning that there is an 8% decrease regarding the best MAP obtained last year and in both years experiments were quite similar. This decrease is the effect of changing the image collection used to test both systems and a clear dependency among retrieval techniques and image collections used to test those techniques can be concluded. It will be interesting to compare results for other participants using both test collections.

7 Future work

A conclusive evaluation of the functionality of the implemented expansion algorithm must be performed. The analysis of obtained results has been started but not still concluded by the time of writing this report. Some failures in the code have been already detected and corrected. The final goal is to include results of experiments run with the reviewed expansion algorithm and compared with the actual ones.

Future work in this image retrieval task will try to exploit semantic information obtained from syntactic analysis and from external resources. Text captions present in the IAPR collection are formed by nominal and prepositional phrases that could be analysed to extract relations among concepts represented by the headers of phrases. Some works in this line will be tested in future campaigns.

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 by the Madrid's R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for Madrid Community), S-0505/TIC/000267.

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, Sara Lana-Serrano, Ángel Martínez-González, Antonio Moreno-Sandoval and César de Pablo-Sánchez.

References

[1] Charniak, Eugene. A Maximum-Entropy-Inspired Parser. In Proceedings of NAACL-2000, 2000.

[2] de Pablo, C.; González-Ledesma, A.; Martínez-Fernández, J. L.; Guirao, J.M.; Martínez, P.; and Moreno, A. MIRACLE's Cross-Lingual Question Answering Experiments with Spanish as a Target Language. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[3] Goñi-Menoyo, J.M.; González-Cristóbal, J.C.; and Villena-Román, J. MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[4] IAPR TC-12 photographic collection: On line http://ir.shef.ac.uk/cloughie/papers/muscle-imageclef2005.pdf [Visited 18/07/2006].

[5] ImageCLEF 2006: On line http://ir.shef.ac.uk/imageclef/2006 [Visited 18/07/2006].

[6] Lana-Serrano, S.; Goñi-Menoyo, J.M.; and González-Cristóbal, J.C. MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[7] LingPipe (Java libraries for the linguistic analysis of human language): On line http://www.alias-i.com/lingpipe/ [Visited 18/07/2006].

[8] Lucene: On line http://lucene.apache.org/ [Visited 18/07/2006].

[9] Martínez-Fernández, J.L.; Villena-Román, J.; García-Serrano, A.M.; and González-Cristóbal, J.C. Combining Textual and Visual Features for Image Retrieval. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[10] 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. Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers (Carol Peters, Paul Clough, Julio Gonzalo, et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 699-708. Springer, 2005.

[11] 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.

[12] Martínez-González, A.; Martínez-Fernández, J. L.; de Pablo-Sánchez, C.; Villena-Román, J. Jiménez-Cuadrado, L.; Martínez, P.; and González-Cristóbal, J.C. MIRACLE at WebCLEF 2005: Combining Web Specific and Linguistic Information. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[13] Montoyo, A. Desambiguación léxica mediante marcas de especificidad, Phd. Thesis, supervised by Manuel Palomar and German Rigau, Universidad de Alicante, 2002

[14] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 18/07/2006].

[15] Villena-Román, J.; Goñi-Menoyo, J.M.; González-Cristóbal, J.C.; and Martínez-Fernández, J.L. MIRACLE Retrieval Experiments with East Asian Languages. Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pp. 138-144. Tokyo, Japan, 2005.

[16] Villena-Román, J.; Crespo-García, R.M.; and González-Cristóbal, J.C. Effect of Connective Functions in Interactive Image Retrieval. Accessing Multilingual Information Repositories: 6th Workshop of the Cross Language Evaluation Forum 2005, CLEF 2005, Vienna, Austria, Revised Selected Papers (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 4022, Springer (to appear).

[17] Villena-Román, J.; González-Cristóbal, J.C.; Goñi-Menoyo, J.M.; and Martínez Fernández, J.L. MIRACLE's Naive Approach to Medical Images Annotation. Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 2005.

[18] Xapian: an Open Source Probabilistic Information Retrieval library. On line http://www.xapian.org [Visited 18/07/2006].