MIRACLE at ImageCLEF 2004

J.L. Martínez-Fernández[1,4], Ana García Serrano[2], Julio Villena[3,4], Víctor David Méndez Sáenz[2], Santiago González Tortosa[2], Michelangelo Castagnone[2] & Javier Alonso[4]

[1] Advaced Databases Group, Computer Science Department, Universidad Carlos III de Madrid,
Avda. Universidad 30, 28911 Leganés, Madrid, Spain
joseluis.martinez@uc3m.es
[2] Artificial Intelligence Department, Universidad Politécnica de Madrid.
Campus de Montegancedo s/n, Boadilla del Monte 28660, Spain
{agarcia, vmendez, sgonzalez, mcastagnone}@isys.dia.fi.upm.es
[3] Department of Telematic Engineering, Universidad Carlos III de Madrid,
Avda. Universidad 30, 28911 Leganés, Madrid, Spain
jvillena@it.uc3m.es
[4] DAEDALUS – Data, Decisiond and Language, S.A.
Centro de Empresas “La Arboleda”, Ctra. N-III km. 7,300 Madrid 28031, Spain
{jalonso, jmartinez, jvillena}@daedalus.es

Abstract. The second participation of the MIRACLE (Multilingual Information RetrievAl for the CLEf campaign) research group in the ImageCLEF task is described in this paper. New techniques, devoted to the combination of linguistic and statistical language processing methods, have been tested, continuing with the experiments carried out in last year

1. Introduction

This is the second time for the MIRACLE (Multilingual Information RetrievAl for the CLEf campaign) research group as a participant in the Image CLEF task. The work presented in this paper is the continuation of the experiments carried out in CLEF 2003. Some new techniques, like the inclusion of linguistic information for monolingual English tasks or the application of EuroWordnet as a translation and query expansion tool, have been developed and tested.

In Image CLEF task, the objective is to deal with textual descriptions of pictures and the corresponding image files. This kind of texts has some specific characteristics, like size and structure of descriptions, making them different from texts used in cross-language tracks. As stated in last year [8], the main focus of the MIRACLE team is to find the way to apply linguistic knowledge to improve the Information Retrieval task. Therefore, for this CLEF call English texts have been treated using tools like the Brill tagger [2], a linguistic parser, a proper noun extraction module and WordNet [4] to include semantic information.

On the other side, this year the MIRACLE team has made a first attempt in analyzing the content of supplied images. For this purpose, GIFT 0.1.9 [6], a public package devoted to image processing has been used. This software can be installed as a server, and some adapted clients based on Viper [6] have been used. Although different search algorithms can be adapted to this tool as plugins, in these experiments the provided *separate normalisation* algorithm has been used.

Image CLEF 2004 offered three different tasks: an adhoc bilingual retrieval task, where images are accompanied by english captions, a medical retrieval task, where a set of scan, x-ray, pictures and short textual descriptions about medical diagnosis are provided, and a user centered search task, where the main goal is to take into account user interaction in the retrieval process. MIRACLE team has taken part in the first two tasks, the first one paying more attention to textual descriptions and the second one to test the content based image indexing and searching tool previously mentioned. As a result, 45 runs have been submitted for both tasks, and a great human effort has been set for this CLEF track.

2. Adhoc Retrieval Task

Figure 1 shows a graphic representation of the different processes followed in the retrieval process according to considered languages.

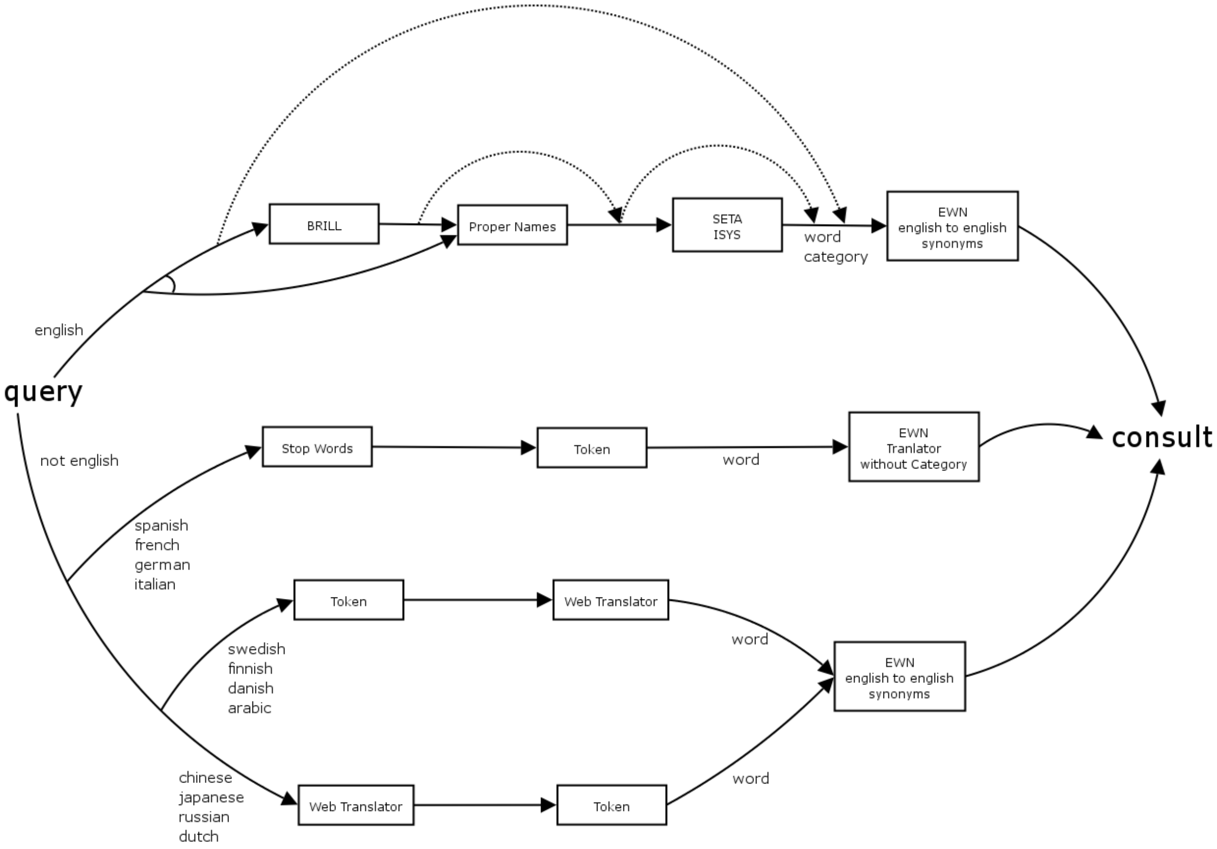

Figure 1. Query processing applied for the Adhoc Retrieval task

As can be seen in the figure, different tools have been used to process english queries:

- **BRILL**: A tagger, based on Brill's work [2], can be used to attach a morphosyntactic tag to each word.
- **Proper Names**: A Proper Name detection module can be applied at the output of the Brill tagger, although plain text can also be used as input to this module.
- **SETA Module**: In a final step, the text can be divided in sentences and their constituent phrases can be extracted using this module. Thus, this module is a linguistic parser for english, implemented using prolog.
- **EWN synonym**: This box represents a subsystem used to extract the corresponding WordNet synonyms for a given word. So, a query expansion is implemented based on semantic information contained in WordNet database. Optionally, the linguistic category of a given word is used when the semantic expansion is performed. For example, if a word acting as a name is going to be expanded, only synonyms of the given word that can act as a name are considered.

The rest of languages have been treated using EuroWordNet, where available, or translation tools, like Systran [1] or Translation Experts [7]. These experiments where devoted to test the quality of EuroWordNet when used in translation tasks and as a synonym expansion tool. For translation purposes, the inter-lingual index (ILI) supplied with EuroWordNet has been applied. Again, it is possible to consider the linguistic category of the word when asking for its translations. So, if a name is going to be translated, only words that can act as a name in the target language are taken into account.

Monolingual English experiments

Figure 1 shows how given queries are processed prior to the search process. Of course, equivalent treatments must be applied in the indexing process. For the English language, Table 1 shows the four different index databases considered and the treatments executed to get each one. In all of them, some common tasks have been applied, like obtaining the stem of the word.

For the rest of languages, only the baseline index database has been used to search translated queries.

Index/Tasks	Tokenize	Filter Stopwords	Brill Tagger	Filter Nouns	Stem ming	Filter Proper Nouns
DB1 - Baseline	√	√	×	×	×	×
DB2 - Only Nouns	×	×	√	√	×	×
DB3 - Proper Names + Baseline	√	√	×	×	√	√
DB4 - Proper Names + Nouns	×	×	√	√	√	√

Table 1. Index databases used in the Adhoc Retrieval task

Table 2 summarizes the different ways applied for query processing, the name given to each experiment and the index database used to perform the search process.

	Monolingual English Experiments		
	Query Process	**Database Searched**	**Run Name**
Baseline	Topic Words	**DB 1**	**mirobaseen**
	Topic Words + Synonyms	**DB 1**	**mirosbaseen**
Only Nouns	Nouns	**DB 2**	**mironounen**
	Nouns + Synonyms without category	**DB 2**	**mirosnounen**
	Nouns + Synonyms with category	**DB 2**	**miroscnounen**
Baseline + Proper Names	Topic Words + Proper Names	**DB 3**	**miroppbaseen**
	Topic Words +Synonyms + Proper Names	**DB 3**	**mirosppbaseen**
Nouns + Proper Names	Nouns + Proper Names	**DB 4**	**miroppnounen**
	Nouns + Synonyms without category + Proper Names	**DB 4**	**mirosppnounen**
	Nouns + Synonyms with category + Proper Names	**DB 4**	**miroscppnounen**
SETA	Topic and Narration Words	**DB 3**	**mirorppbaseen**
	Topic and Narration Words + Synonyms with category	**DB 3**	**mirorscppbaseen**

Table 2. Run definitions for the Adhoc Retrieval task

In this table, 'Topic Words' means that all simple words (excluding stopwords) are used to search the corresponding index database. 'Synonyms' means that all synonyms for a word found in WordNet are used to expand the query, without any refinement. 'Nouns' stands for the situation where the query text is tagged and only words acting as nouns are selected as part of the final query. 'Proper Names' is used to mark that only recognized proper names in the text are used as part of the query. 'Synonyms with category' is used to distinguish the process in which not all the synonyms of a words are taken into account, but only those synonyms that can act with the same category than the initial word are included in the query. Finally, in the last two experiments included in the table, the narrative of the query (only available for the english queries) is used as the input to the SETA module, in charge of parsing the text and getting a more precise category for the word.

Monolingual English Results

Table 3 shows average precision figures (MAP column) and the position obtained for each defined run (described in the previous section).

Run Name	MAP	Rank
mirobaseen	0,5865	1
mirosbaseen	0,5623	4
miroppbaseen	0,5609	6
mirosppbaseen	0,5388	8
miroppnounen	0,3384	87
mirosnounen	0,3383	88
mirorppbaseen	0,3366	90
mirosppnounen	0,3337	92
mirorscppbaseen	0,2703	112
miroscppnounen	0,2568	116
mironounen	0,2525	119
miroscnounen	0,2461	120

Table 3. Average precision results for monolingual english experiments

Regarding these results, it is important to highlight some points: first of all, the basic experiment (taken as the baseline) produces the best results. Very different results are obtained from the fourth result in advance and again a gap in the average precision can be seen from the eighth result in advance. These differences in precision show that, when all words are used in the characterization of the textual captions results are better and the inclusion of more linguistic information (like proper nouns or synonyms) does not lead to an improvement. On the other side, if only common or proper nouns are used to represent the documents there is a loss in precision, perhaps due to the fewer number of words used for document characterization. Also, its worth mentioning that the experiment using all linguistic information that available tools can extract is among the worse ones

Bilingual Experiments

For the bilingual experiments two different approaches, depending on available information, have been applied. These two approaches are:

- A EuroWordNet based approach, where information contained in the ILI index provided by EuroWordNet is used to translate the original query. This approach has been applied for the following languages: Spanish, German, French and Italian.
- A translator based approach, where online translation tools, in particular Systran and Translation Experts tools, have been applied to translate the queries from the initial language to the target language (English in Image CLEF tasks).

In both approaches, the index database used corresponds to the baseline, i.e., the one where all words (excluding stopwords) are considered as indexes. Table 4 summarizes the features of the experiments defined for this multilingual task.

	Multilingual Experiments				
	Tokenize	Remove Stopwords	Translation method	Index Database	Run Names
EWN Languages	√	√	EuroWordNet translation module without categories	DB 1	mirowbaseit mirowbasefr mirowbasees mirowbaseesc mirowbasege
Russian, Japanese and Chinese	√	×	Automatic translation using web translator: BabelFish (*http://babelfish.altavista.com*)	DB 1	mirobaseru mirobaseja mirobasezh
Finnish, Swedish and Danish	×	×	Automatic translation using web translator: TransExp (*http://www.tranexp.com*)	DB 1	mirobasedu mirobasesw mirobaseda

Table 4. Description of Multilingual Experiments for the Adhoc Retrieval task

Bilingual Results

Table 5 shows average precision figures obtained for the multilingual experiments defined in the previous section.

Run Name	MAP	%Monolingual	Rank
mirobaseru	0,3866	65,93	73
mirobasedu	0,3807	64,91	76
mirobasesw	0,3043	51,89	99
mirowbaseit	0,2857	48,72	106
mirobaseda	0,2799	47,72	107
mirowbasees	0,2687	45,82	113
mirowbaseesc	0,2615	44,59	114
mirowbasege	0,2455	41,87	122
mirobaseja	0,2358	40,21	124
mirowbasefr	0,2188	37,31	127
mirobasezh	0,1777	30,30	135
mirobasefi	0,17	28,99	141

Table 5. Average precision for Multilingual Adhoc Retrieval experiments

According to these results, one important fact to mention is the loss of precision, taking into account the best monolingual experiment. As can be seen, a decrease of 34,07% in precision, marking again the importance of the quality of the translators used in multilingual environments. Cases where EuroWordNet has been used as a translation tool can be compared with CLEF 2003 obtained results [8] and an important decrease in precision can be noticed. Last year bilingual experiments with French, German, Italian and Spanish where around 40% average precision, while this year average precision for these languages is around 30%. It is also worth mentioning that other participants, according to official results, have obtained only a decrease of 10% in precision for some bilingual tasks, so, in our situation, there is room for improvement.

Mixing text based retrieval with content based image retrieval (CBIR) for the Adhoc task

This year, the MIRACLE team has made a first step in image content retrieval. This first step has led to the definition of experiments where content based image retrieval (CBIR) is applied. This is the case of the adhoc retrieval task, where some runs mixing results obtained using textual search and CBIR search have been submitted. The CBIR subsystem used for this experiment is based on GIFT 0.1.9 and will be described in the next section. The text retrieval subsystem is the one used in text based experiments, although for initial test and tuning of the overall system, last year data and text search systems have been used.

The process of mixing textual and image results begins taking the list with the images returned by the text search subsystem and their relevance figures and building a query for the CBIR subsystem. The content search is performed and a new search is performed considering the 5 first elements returned. Finally, results obtained with this last relevance feedback approach are combined with the original results list returned by the textual search subsystem. The expression used to combine these partial lists is:

$$\begin{cases} \sqrt[k]{REL\_VIS^{weight\_vis} \times REL\_TXT^{weight\_txt}}\ , & \text{for elements in both lists} \\ & \text{and k = weight\_vis + weight\_txt} \\ \text{factor\_vis,} & \text{for elements appearing only in the list obtained with the CBIR subsystem} \\ \text{factor\_txt,} & \text{for elements appearing only in the list obtained with the textual search subsystem} \end{cases}$$

In this expression, *REL_VIS* and *REL_TXT* are the relevance value returned by the CBIR subsystem and the text search subsystem respectively. *factor_vis*, *factor_txt*, *weight_vis* and *weight_txt* are parameters to be defined and can be used to adjust the overall system according to obtained results, for example, giving more importance to textual results or CBIR results.

Results of the text and CBIR mixing experiments

Two sets of experiments have been done. Results for the first set are included in Table 6, where the initial set of text search results have been some of the experiments defined in Table 2.

Run Name	MAP	Rank	Initial Text search Experiment
enenrunexp1	0,5838	2	mirobaseen
enenrunexp7	0,5339	9	mirosppbaseen
enenrunexp4	0,3373	89	mirosnounen
enenrunexp10	0,2533	118	miroscppnounen

Table 6. Average Precision values for text and CBIR mixing experiments

Comparing to Table 3, these results are very close (and always below) of the ones where only textual search is applied. This can be due to the chosen configuration of the combination algorithm. More tests should be made to extract a valid conclusion.

Some other experiments where executed using a different textual search subsystem. Obtained results for these experiments (Table 7 and Table 8) have been always worse than the previously mentioned ones. One of these sets of experiments, Table 8, is a bilingual one with English as a target language and Spanish as the initial language.

Run Name	MAP	Rank
enenrun8	0,4173	52
enenrun7	0,3389	86
enenrun1	0,3362	91
enenrun4	0,0737	186

Table 7. Average Precison values using a different text retrieval system

Run Name	MAP	%Monolingual	Rank
esenrun8ok	0,1226	20,91	164
esenrun2ok	0,1206	20,57	166
esenrun7ok	0,0787	13,42	183
esenrun1ok	0,0783	13,35	184

Table 8. Average Precision values using a different text retrieval system (bilingual Spanish)

3. Medical Retrieval Task

This year ImageCLEF organizers have defined a new task where the main focus is image content based retrieval. For this purpose a set of medical images, including scans, x-ray images and photographs of different illness has been made available to ImageCLEF participants.

The CBIR system used has been GIFT 0.1.9 [6] developed under GNU licence which allows query by example, using an image as a starting point for the search process, and implements relevance feedback methods. This software has been developed by the Vision Group at the CUI of the University of Geneva.

Although the first step in the search process for this task must involve an image, textual descriptions of the medical cases have been used to try to improve retrieval results.

The search process can be divided in the following steps:

1. The initial query, formed by one image, is introduced in the CBIR system to obtain a set of images to define the query.
2. The CBIR system returns a list of images along with the corresponding relevance values. The number of images used in the search process is called relevance threshold and constitutes a system configuration parameter.
3. Previous steps have produced a valid query which is introduced in the overall system. The complete system is formed by a textual subsystem and a CBIR subsystem. In a first step both subsystemas are used to perform the search process.
4. Partial results lists are combined using an intersection operator: images not appearing in both partial lists are dropped. Two special parameters make it possible to consider textual results more important than CBIR ones or vice versa.
5. The previous step produces a unique results list that is again introduced in the CBIR subsystem. The new results list obtained is again combined, applying the intersection operator, with the output of the textual subsystem.

The overall process is depicted in Figure 2. The expression used to obtain a unique relevance value according to the partial results lists produced by textual and CBIR subsystems is:

$$\text{Rel} = \begin{cases} \sqrt[k]{REL\_VIS^{weight\_vis} \times REL\_TXT^{weight\_txt}}, & \text{for elements in both lists and k = weight\_vis + weight\_txt} \\ 0, & \text{for elements appearing only in one partial list} \end{cases}$$

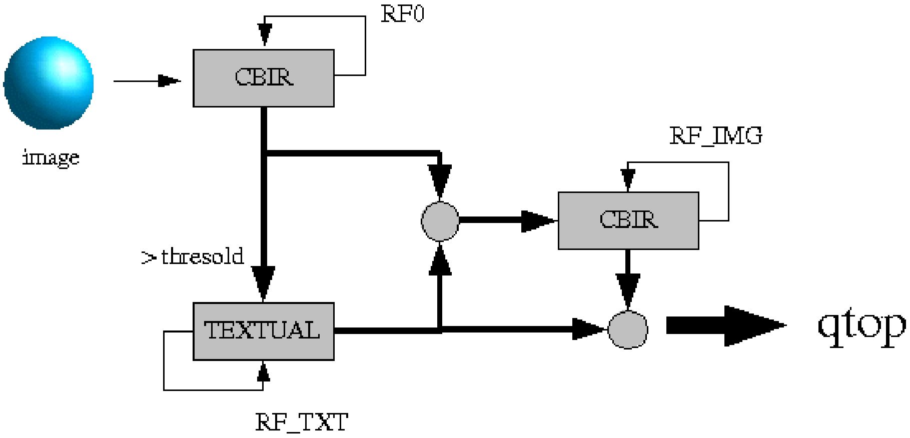

Figure 2. Text and CBIR subsystems combination model

Some other models, based on different ways to combine the output of the textual and CBIR subsystems, have been tested but the one described here has produced the best results. Four different runs have been defined according to different values for the configuration parameters defined for the overall system. These parameters include: the minimum threshold to build the initial query, the number of results used for relevance feedback and the weights given to textual and image results.

Medical Retrieval Task Results

Average precision figures obtained for the submitted experiments are included in Table 9. As can be seen, the difference in precision among the first and the last run is around 2%, not enough to extract some conclusions about which method (or configuration parameters set) is the best. On the other hand, according to the obtained rank for these runs, there is still room for improvement in this task, perhaps testing new configurations or new values for defined parameters or taking the most of textual descriptions related to each medical case.

Run Name	MAP	Rank
enid1run	0.1798	32
enid3run	0.1752	33
enid0run	0.1650	34
enid2run	0.1542	35

Table 9. Average Precision values for Medical Retrieval experiments

6. Conclusions

This is the second year for the MIRACLE team taking part in the CLEF campaign and in the ImageCLEF track in particular. The main goal pursued this year was to continue with the research in finding a right combination of linguistic and statistical methods to improve the Information Retrieval process. MIRACLE group is also very interested in the field of multimedia retrieval so, the content based image retrieval task defined this year as part of the ImageCLEF track was a great opportunity to take a first step in the field. From out point of view, obtained results for the adhoc retrieval task are very good. Average precision values for the monolingual english task are a little bit better than the ones obtained last year, pointing that is difficult to improve results for this task. Perhaps the best performance figures that can be obtained with actual technology have been reached. On the contrary, bilingual tasks, in the way we have developed them, can be improved.

A mention apart must be made for the content based image retrieval task, where obtained results are not as good as for the textual task. This fact drives us to increase efforts devoted to this kind of retrieval for the following campaigns.

7. Acknowledgements

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

References

[1] “*Altavista’s Babel Fish Translation Service*”, http://babelfish.altavista.com/, last accessed 12.08.2004

[2] Brill E., "*Some Advances in Transformation Based Part of Speech Tagging*", proceedings of the Twelfth National Conference on Artificial Intelligence, 1994

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

[4] G.A. Miller. “*WordNet: A lexical database for English*”. Communications of the ACM, 38(11):39—41,1995

[5] “The Porter Stemming Algorithm” page maintained by Martin Porter. http://www.tartarus.org/~martin/PorterStemmer/, last accessed 12.08.2004

[6] "*The GNU Image-Finding Tool. GIFT 0.1.9*", http://www.gnu.org/software/gift/, last accessed 12.08.2004

[7] “*Translation Experts*”, http://www.transexp.com, last accessed 12.08.2004

[8] Villena J., Martinez-Fernandez J.L., Fombella J., García-Serrano A., Ruíz-Cristina A., Martínez P., Goñi J.M., González-Cristóbal J.C., “Image Retrieval: the MIRACLE Approach”, In CLEF 2003 Proceedings, Springer-Verlag, to appear, 2003