Boolean Operators in Interactive Search

Julio Villena-Román[1,3], Raquel M. Crespo-García[1]
José Carlos González-Cristóbal[2, 3]

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

jvillena@daedalus.es, rcrespo@it.uc3m.es
jgonzalez@dit.upm.es

Abstract

This paper presents the participation of the MIRACLE team at the ImageCLEF interactive search task. Basically, queries consisting on several terms can be processed combining their words using either an AND function or an OR function. The AND approach forces the user to use precise vocabulary and query terms must exactly match the terms in the index for the target to be found. However, this is quite difficult to integrate in cross-lingual systems with automatic translation, as many terms can turn out to be ambiguous and accept different translation options. The OR approach allows less precise vocabulary and more ambiguous translations, and also relevance feedback can be used to achieve the search goals. From the user's point of view, the AND approach seems to be more intuitive because the system responses can be made as precise as wanted, simply by adding more words to the query. On the other hand, with the OR approach, the more terms are included in the query, the more images are probably recovered. In a cross language scenario, the AND approach can prove to be difficult for non-native speakers, particularly in specialised tasks which require domain-specific vocabulary as the one modelled in our experiment. In such conditions, the OR approach assisted by automatic translation can be a more helpful choice. Our idea is to compare if the search success rate is similar for both approaches.

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, iCLEF, image retrieval, vocabulary, precision, Boolean operators, query.

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 evaluating and exchanging state of the art knowledge in the field of cross-language information retrieval systems, held annually since 2000. ImageCLEF provides a realistic benchmark for image retrieval systems, either based on image visual characteristics, semantic textual information aggregated to them, or hybrid systems combining both approaches. As any interactive information retrieval system, user interaction is a key factor in the systems effectiveness. Feedback received, as well as search strategy, has proved to be determinant to help the user to achieve his/her goals. A user-centered task (interactive) is thus offered as a part of the ImageCLEF benchmark.

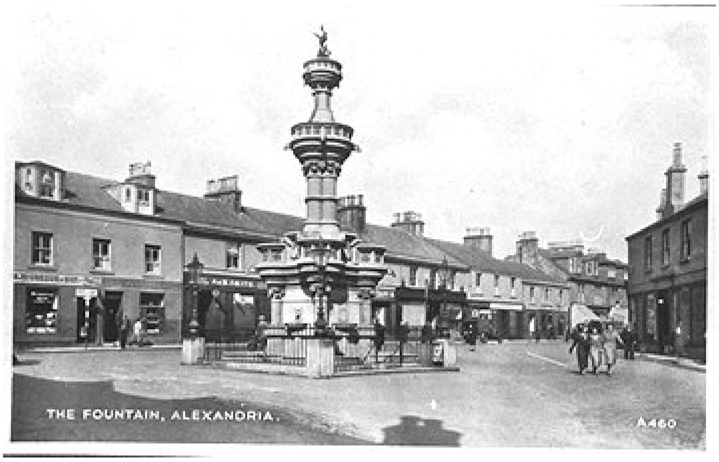
Images are inherently language independent and thus image retrieval can often be seen as a language-independent task. Results (images) can be presented visually, with no text. Even the queries can be done visually, by searching images similar to another one. However, searching images using a text-based query interface, based on image descriptions or metadata is a common scenario. Also, the presentation of the resulting images can be complemented including their corresponding descriptions or using their metadata. Thus, image retrieval integrates with cross-language information retrieval, as users' native language (or even languages) can differ from the language used for labelling the image collection.

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 [4],[5],[6],[7],[8],[12],[13]. As well as bilingual, monolingual and cross lingual tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and GeoCLEF tracks.

In this paper, the results of the experiments deployed by the MIRACLE group on the ImageCLEF interactive search task are presented. First, the problem of cross-language image retrieval is introduced in the present Section 1, together with CLEF context and MIRACLE work. Section 2 contains the task objectives and the description of the database. Experiments are described in Section 3. In Section 4, a description of the MIRACLE toolbox developed for testing the hypothesis is provided. Methodology and results are presented in Section 5 and 6, followed by conclusions in Section 7.

2 Task Description

The St Andrews image collection is used for the ImageCLEF interactive search task. This collection contains 28,133 photographs from the St Andrews University Library photographic collection [10], one of the largest (over 300,000 images) and most important collections of historic photography in Scotland. Photos are primarily historic in nature from areas in and around Scotland; although pictures of other locations also exist. The majority of images (82%) are in black and white, although colour images are also present in the collection. All images have an accompanying textual description consisting of 8 distinct fields. All captions are written in British English, although the language also contains colloquial expressions. An example is shown in Figure 1 .

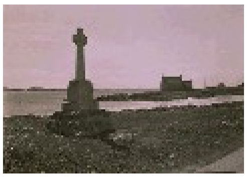
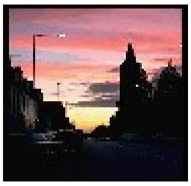
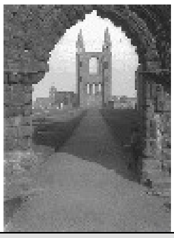
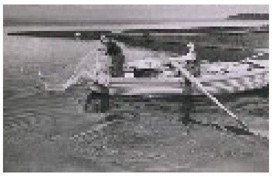
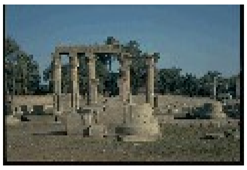
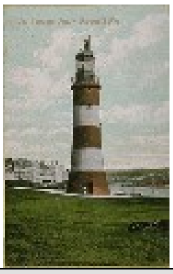
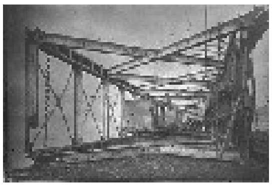
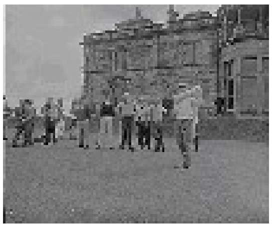
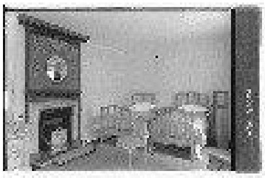
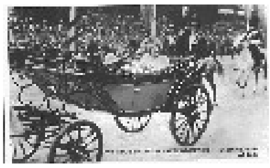
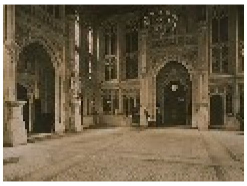
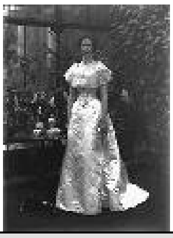
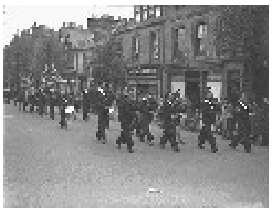
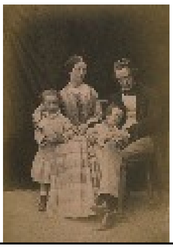
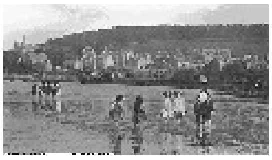
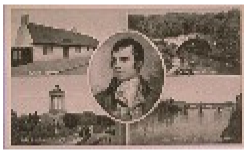


Record ID	JV-A.000460
Short Title	The Fountain, Alexandria.
Long Title	Alexandria. The Fountain.
Description	Street junction with large ornate fountain with columns, surrounded by rails and lamp posts at corners; houses and shops.
Date	Registered 17 July 1934
Photographer	J Valentine & Co
Location	Dunbartonshire, Scotland
Notes	JV-A460 jf/mb
Categories	[columns unclassified][street lamps - ornate][electric street lighting][shepherds & shepherdesses][streetscapes][shops]

Figure 1. Example of an image caption in the St. Andrews collection

Given an image (not including the caption) from the St Andrews collection, the goal for the searcher is to find the same image again using a cross-language image retrieval system. This models the situation in which a user searches with a specific image in mind (perhaps they have seen it before) but without knowing key information thereby requiring them to describe the image instead, e.g. searches for a familiar painting whose title and painter are unknown. The 16 images used as topics for each search task are shown in next table.

Table 1: Topic images

Topic 1	Topic 2	Topic 3	Topic 4
			
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
			
Topic 9	**Topic 10**	**Topic 11**	**Topic 12**
			
Topic 13	**Topic 14**	**Topic 15**	**Topic 16**
			

3 Experiments

Queries consisting on several terms can be processed combining their words using either an AND function or an OR function [12][13]. The AND approach forces the user to use precise vocabulary. Query terms must be exactly included in the index for the image to be found. Also, the system responses can be made as precise as wanted, simply by adding more words to the query. However, this is quite difficult to integrate in cross-lingual systems with automatic translation, as many terms may turn out to be ambiguous and accept different translation options. A more “fuzzy” and noisy search results from the OR approach. However, it allows less precise vocabulary and more ambiguous translations. In this case, relevance feedback can be used to achieve the search goals instead of image filtering.

From the user’s point of view, the AND approach seems to be more intuitive. If the result set is too large, the search can be refined by including more search terms. If it is too reduced, that is, too many images have been filtered out from the solution, it can be broadened just reducing the requirements included in the query. That is, the more specific the query is, the more specific result set is generated. An immediate sense of control results from this approach, as the solution set reduces its size as the user approaches to the goal.

On the other hand, the OR approach seems to be less effective from the user’s point of view. The more terms are included in the query, the more images are probably recovered. Although relevance order is probably more accurate, the user perceives a more generalized, less precise result, as the result set includes more images.

In a cross language scenario, the AND approach can prove to be difficult for non-native speakers, particularly in specialised tasks which require domain-specific vocabulary as the one modelled in our experiment. In such conditions, the OR approach assisted by automatic translation can be a more helpful choice.

Our idea is to compare whether it is better to make AND English queries (which have to be precise and use the exact vocabulary, which maybe difficult for a specialised task like this) or to use OR queries in Spanish and have the option of relevance feedback (a more "fuzzy" and noisy search but which doesn't require precise vocabulary and exact translations). We are going to study if, in the context of an interactive search task, users prefer either a system with higher precision but smaller result sets (the AND monolingual system) or a system gives more results although less precise (the OR bilingual system). Our objective is to check if the search success rate is similar for both approaches.

4 MIRACLE Toolbox

Two systems have been developed to test and compare each search strategy previously described. System A, **miraML**, implements the AND monolingual approach, whereas System B, **miraCL**, implements the OR bilingual approach. A similar web-based user interface is used for both of them.

Both systems, miraML and miraCL, execute the queries against a common index, built from the collection of images, using Xapian, a publicly available text search engine [14]. Among all the available metadata, only the description of the images is used for generating the index, as it is the field which includes the most useful information for the searches. Natural language processing techniques are applied before indexing. An adhoc language-specific parser for English is used to identify different classes of alphanumerical tokens such as dates, proper nouns, acronyms, etc., as well as recognising common compound words. Image descriptions are tokenised, stemmed [9] and stop word filtered [11] to improve searching efficiency (Figure 2).

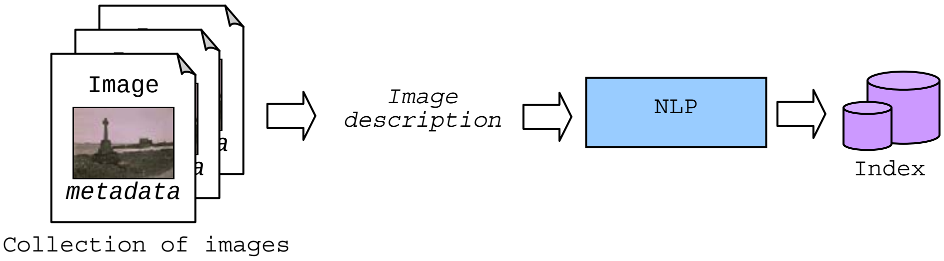


Figure 2: Index generation.

To evaluate the results, both systems keep a log with the user queries and the number of results obtained for each one. Time is also logged, using a timer that is started when the user begins a new topic. The time to execute the queries or to provide the automatic translation is not taken into account for the user time, simply by disabling the timer when performing those operations.

4.1 System A: miraML

miraML is a pure monolingual system. User queries are made in English, the source language of the image collection. As well as for the image descriptions, a stemming process is applied to all the words included in the query. Query stems are then concatenated with the AND operator before executing the search.

Results are displayed combining visual and textual presentation. The result page shows both the images and their associated textual descriptions. Only the 20 most relevant results are presented to the user, with no pagination.

As a help for the user, a Spanish to English translation textbox is provided, which queries FreeTranslation.com [2], Altavista BabelFish [1], Google Translation [3] and I2E programme (included in Debian Linux distribution). If several options exist for a given term, the user must select the appropriate translation.

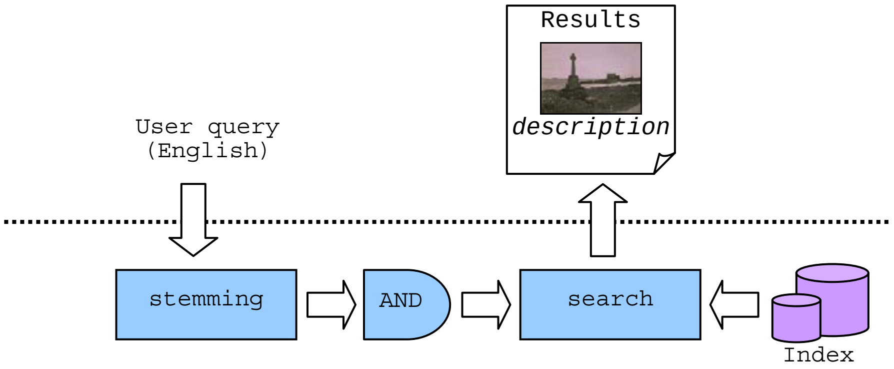


Figure 3: miraML query processing.

4.2 System B: miraCL

miraCL is a cross-lingual (bilingual) Spanish to English image retrieval system. Queries are expressed in Spanish, in this case the native language of the users, and automatically translated into English. A stemming process is applied to the English terms resulting from the translation phase. As the user can include several terms, and each of them can accept different translations, English resulting stems are concatenated with the OR operator and finally executed. Figure 4 sketches the process.

Also, the user may use relevance feedback and ask for similar images to a given list of images. The system builds a new query concatenating the first 25 more relevant keywords of each image in the list.

The same as system A (miraML), only 20 best results are presented to the user. However, in contrast, the result page only shows the images with their ID and relevance, with no descriptions, because no English text should be shown to the users.

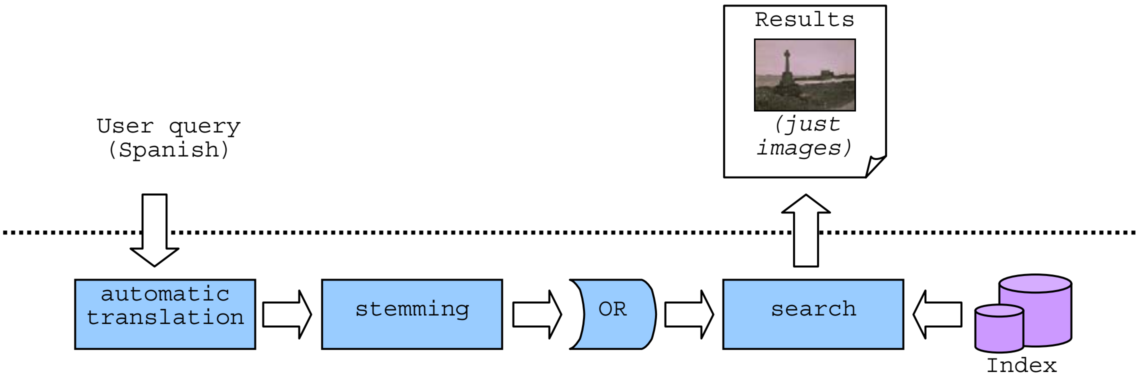


Figure 4: miraCL query processing.

5 Methodology

Experiments have be designed and deployed following ImageCLEF guidelines. Eight persons have participated. A brief tutorial was given to the participants about the systems, and they also performed some test search during a training phase, to learn to use them. Each one was asked to search 16 topics (images), half of them with each system. In order to distinguish the influence of the different factors involved, mainly users, topics and systems, a latin-square design has been applied to assign which system should use each participant to search each topic, as shown in Table 2. However, some users didn't understand the instructions correctly (or the instructions weren't correctly given) and didn't use the correct system in their tasks, so the latin-square assignment was not strictly followed.

Table 2: Latin-square assignment of searching system to each user and topic.

Topic Searcher	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	A	A	A	A	B	B	B	B	A	A	A	A	B	B	B	B
2	B	B	B	B	A	A	A	A	B	B	B	B	A	A	A	A
3	A	A	A	A	B	B	B	B	A	A	A	A	B	B	B	B
4	A	A	A	A	B	B	B	B	A	A	A	A	B	B	B	B
5	A	B	B	A	B	A	A	B	A	B	B	A	B	A	A	B
6	B	A	A	B	A	B	B	A	B	A	A	B	A	B	B	A
7	A	B	B	A	B	A	A	B	A	B	B	A	B	A	A	B
8	B	A	A	B	A	B	B	A	B	A	A	B	A	B	B	A

The user profile is rather similar for all participants. Native language is Spanish for all persons involved, but all of them reported a quite fluent level in English, the original language of the captions for the images database. A similar percentage of male (5) and female (3) was included in the population.

6 Results

Table 3 shows that both systems are similar in the success rate (finding the target image in 5 minutes).

Table 3: Success rate for each system.

System	Found	Not found	Success rate
A (miraML)	44	20	68.75%
B (miraCL)	42	22	65.63%

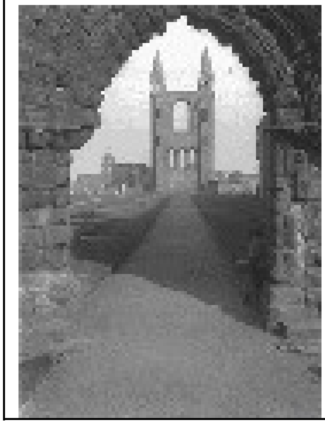
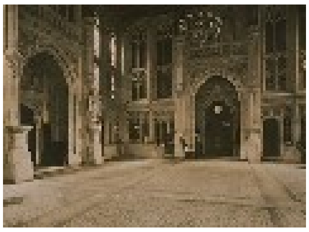
As shown in Table 4, this rate is rather independent of both the searcher and the system but depends more on the topic. Some topics were specially difficult to find, in most cases, due to the poor descriptions. For example, only one user managed to find topics 3 and 11. In the case of topic 3 (see Table 5), all users focused on “stone arch”, “path” or “woman sitting”, features which appear at the foreground, but the image caption only included a few references to St Andrews cathedral. The description of topic 11, the hall of the House of Commons, only mentioned the two arched doors and the small clock over one of them, but no references at all to the statues, lamp, columns or floor.

Table 4: Success rate per topic and system.

Topic	System A (1)	System B (1)	Total (1)	System A Success rate	System B Success rate	Total Success rate
1	5/0	3/0	8/0	100,0%	100,0%	100,0%
2	5/0	2/1	7/1	100,0%	66,7%	87,5%
3	1/4	0/3	1/7	20,0%	0,0%	12,5%
4	5/0	3/0	8/0	100,0%	100,0%	100,0%
5	1/2	1/4	2/6	50,0%	20,0%	25,0%
6	3/0	4/1	7/1	100,0%	80,0%	87,5%
7	2/1	4/1	6/2	66,7%	80,0%	75,0%
8	2/1	3/2	5/3	66,7%	60,0%	62,5%
9	5/0	3/0	8/0	100,0%	100,0%	100,0%
10	3/2	1/2	4/4	60,0%	50,0%	50,0%
11	0/5	1/2	1/7	0,0%	50,0%	12,5%
12	3/2	3/0	6/2	60,0%	100,0%	75,0%
13	3/0	4/1	7/1	100,0%	80,0%	87,5%
14	3/0	5/0	8/0	100,0%	100,0%	100,0%
15	1/2	2/3	3/5	50,0%	40,0%	37,5%
16	2/1	3/2	5/3	66,7	60,0%	62,5%

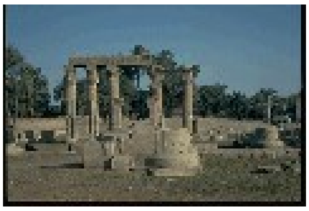
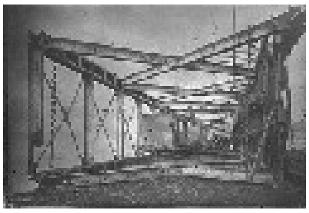
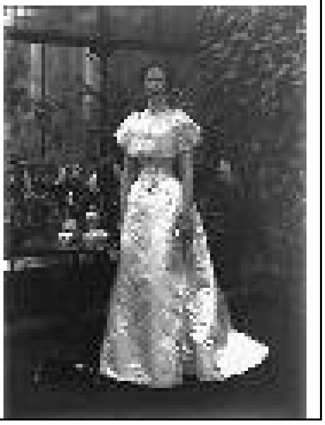
(1) Found/Not Found

Table 5: Most difficult topics.

Topic 3	Topic 11
	

Only for a few topics, one of the system turned to be more helpful to the users than the other. In the case of topic 5 (Table 6), System A was better that System B (50% against 20% in success rate), because of the poor description of the image, which required a very precise query ("ruins temple") instead of a more general query with "columns", or "monument". For topic 12, most users failed to give a precise description of the photograph using "specialised" vocabulary such as "necklace" or "satin dress". In this case, System B was more adequate because the relevance feedback with another photograph of Lady Gilmour provided the target image. The same happened with topic 7, in which most users didn't understand the photograph or didn't know the terms to describe it (e.g., "iron girder"). System B allowed to use feedback with similar images from the Tay Bridge disaster to find the topic image.

Table 6: System-oriented topics.

Topic 5	Topic 7	Topic 12
		

Next table shows the success and failure rates per user and system.

Table 7: Success and failure rates per user and system.

System		Searcher								% Total
		1	2	3	4	5	6	7	8	
A	%success	75,0%	87,5%	75,0%	50,0%	75,0%	50,0%	75,0%	62,5%	68,75%
	%failures	25,0%	12,5%	25,0%	50,0%	25,0%	50,0%	25,0%	37,5%	31,25%
B	%success	50,0%	62,5%	75,0%	75,0%	37,5%	87,5%	50,0%	87,5%	65,63%
	%failures	50,0%	37,5%	25,0%	25,0%	62,5%	12,5%	50,0%	12,5%	34,38%
Total %success		62,5%	75,0%	75,0%	62,5%	56,3%	68,8%	62,5%	75,0%	67,19%
Total %failures		37,5%	25,0%	25,0%	37,5%	43,8%	31,3%	37,5%	25,0%	32,81%

Regarding the time taken for the queries, Table 8 (a, b) show that there seems to be no difference in both systems, as some users take longer with System A than System B, and vice versa. The average time is similar for both systems.

Table 8a: Average searching time per topic and system (for successful searches, in seconds).

System	Topic																Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
A	97	60	134	34	95	78	212	196	82	168		153	170	93	61	22	102
B	85	119		139	226	127	163	196	33	155	156	113	51	68	181	63	113
Total	92	77	134	73	161	106	179	196	64	165	156	133	102	77	141	47	107

Table 9b: Average searching time per user and system (for successful searches, in seconds).

System	Searcher								Total
	1	2	3	4	5	6	7	8	
A	74	146	101	56	70	130	117	113	102
B	97	80	181	102	101	135	74	100	113
Total	83	118	141	83	80	133	100	105	108

An interesting result is shown in Figure 5. The lower the success rate is for a user, the less average time he/she takes to find the image when he/she successes: it seems that search speed and success rate are inversely dependent. Although statistical significance is arguable with this reduced user set, a possible explanation relies on the search strategy of those users. Their initial impression of the image suggested them a search line, certain terms and certain objects in the picture. The sharpest this initial approach was, the fastest he/she finds the image. But, also, the highest the failure probability was if the expected terms did not appear in the image description. Because the user was the most convinced of his/her approach and do not vary too much the query.

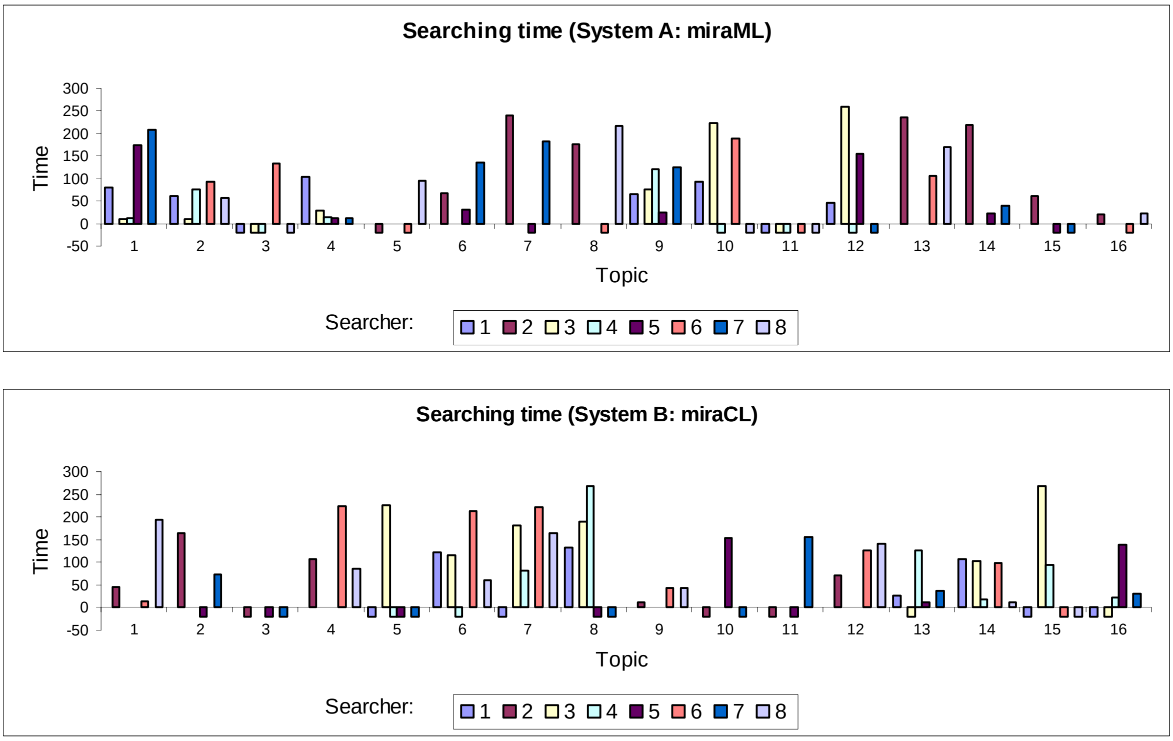


Figure 5 (a, b): Searching time for each system, topic and user combination (negative values stands for failed searches).

Interviews were made after the experiments to know users' subjective opinions. Most users say that System A is "better" than System B because they prefer precise queries as they know the target language, but recognise that, in some occasions, the automatic translation helped them to find the exact word. After the first query with System A, usually users don't think in more terms but scroll the results to add terms that appear in images that are similar to the target or that contain some common features. Apart from descriptions, most users use other photograph attributes such as location (“Bangor Bay”, “St Andrews cathedral”), name of the person (“lady Gilmour”) or a category descriptor (“lighthouses”, “club houses”). Sometimes users start with an introductory query (like “family”), to obtain a first set of results which they use to extract words from the image captions.

Most users complained about descriptions of images 3 and 11, saying that a too specific vocabulary was used or that expected terms weren’t included. Users also complained that finding a given image in time often depended on finding the appropriate keyword included in its description (for example, “necklace” for topic 12). If the user did not know it, or did not realise to include it in the query, he/she could not find the image, though sometimes these keywords did not seem to be the most representative of the picture.

It is also interesting to remark that, after an unsuccessful search, some users continued with the same topic because they did want to find the image. In most cases, users managed to find the topic a short time (<1 minute) after the limit. Due to lack of time, no analysis has been done, but this fact may be taken into account for next year, for example, extending the time limit or considering a strict limit and a less-strict one.

7 Conclusions and Future work

Our experiment compared whether it is better for users to make AND English queries (which have to be precise and use the exact vocabulary) or to use OR queries in Spanish (which are automatically translated by the system) and have the option of relevance feedback (a more "fuzzy" and noisy search but which doesn't require precise vocabulary and exact translations). Results show that there is no significant difference between systems and the search success rate is similar for both approaches.

This conclusion is very promising and will further investigated because of its application to another scenario in which our group is very interested: teaching and/or learning improvement. Our (Spanish) students have to read documents and references which are usually written in English. In many cases this is an actual challenge for them which imposes constrains and limitations in their learning rate and the activities that can be done in class or at lab. This fact may be specifically observed when students have to look for some particular information in API documentation: they refuse to search and prefer to ask other students or the professor, which limits their autonomy and their learning process. Should the conclusions of our experiment at iCLEF be applicable to this scenery, searching in English (System A) or in Spanish (System B) may offer the same searching performance, thus eliminating those constrains and improving the learning process of students.

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.

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

References

[1] BabelFish Altavista. Free text translator. On line http://babelfish.altavista.com [Visited 20/07/2005].

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

[3] Google Translate. Free text translator. On line http://www.google.com/translate_t [Visited 20/07/2005].

[4] 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).

[5] 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.

[6] 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).

[7] 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.

[8] 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.

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

[10] St. Andrews Photographic image collection. On line http://specialcollections.st-and.ac.uk/photcol.htm [Visited 13/07/2005]

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

[12] 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.

[13] 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.

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