Overview of the CLEF 2005 Interactive Track

Julio Gonzalo,* Paul Clough† and Alessandro Vallin‡

Abstract

The CLEF Interactive Track (iCLEF) is devoted to the comparative study of user-inclusive cross-language search strategies. In 2005, we have studied two cross-language search tasks: retrieval of answers and retrieval of annotated images. In both tasks, no further translation or post-processing is needed after performing the tasks to fulfill the information need.

In the interactive Question Answering task, users are asked to find the answer to a number of questions in a foreign-language document collection, and write the answers in their own native language. In the interactive image retrieval task, a picture is shown to the user, and then the user is asked to find the picture in the collection.

This paper summarizes the task design, experimental methodology, and the results obtained by the research groups participating in the track.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.4 [**Information Systems Applications**]: H.4.m Miscellaneous

General Terms

Interactive Information Retrieval, Cross-Language Information Retrieval, Question Answering, Image retrieval

Keywords

information retrieval, interactivity, cross-language,user studies

1 Introduction

In CLEF 2005, user studies have consolidated the two research issues studied in CLEF 2004 as pilot tasks: cross-language question answering and known-item image search.

In the **interactive Question Answering task**, users are asked to find the answer to a number of questions in a foreign-language document collection, and write the answers in their own native language. Subjects must use two interactive search assistants (which are to be compared), pairing questions and systems according to a latin-square design to filter out question and user effects. For this task, we have used a subset of the ad-hoc QA testbed, including questions, collections and evaluation methodology.

In the **interactive image retrieval task**, a picture is shown to the user, and then the user is asked to find the picture in the collection. This was chosen as a realistic task (finding stuff I've seen before) in which visual features could also play an important role (users are given a

*Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, E.T.S.I Informática, Ciudad Universitaria s/n, 28040 Madrid, SPAIN, julio@lsi.uned.es

†Department of Information Studies, Sheffield University, UK, p.d.clough@sheffield.ac.uk

‡ITC-irst, Trento, Italy, vallin@itc.it

picture instead of a written description of what they have to look for). The target data is the St. Andrews' collection (as used in the ad-hoc image CLEF task), in which images are annotated in English with a number of rich metadata descriptions. Again, each participant group was expected to compare two different search assistants, combining users, queries and systems according to a latin-square desing to filter out query and user effects.

The remainder of this paper describes the experimental design and the results obtained by the research groups for each of these tasks.

2 Image Retrieval task

The ImageCLEF interactive search task provides user–centered evaluation of cross–language image retrieval systems. In cross–language image search, the object to be retrieved is an image. This is appealing as a CLIR task because often (depending on the user and query) the object to be retrieved (i.e. the image) can be assumed to be language-independent, i.e. there is no need for further translation when presenting results to the user. This makes a good introductory task to CLIR, requiring only query translation to bridge the language gap between the user's query (source) language, and the language used to annotate the images (target language).

Image retrieval can be purely visual in the case of query–by–example (QBE) which is entirely language–independent, but this assumes the user wants to perform a visual search (e.g. find me images which appear visually similar to the one provided). However, users may also want to search for images starting with text-based queries (e.g. Web image search) requiring that texts are associated with the target image collection. For CLIR, the language of the texts used to annotate the images should not affect retrieval, i.e. a user should be able to query the images in their native language making the target language transparent. Effective cross–language image retrieval will involve both text–based and content–based IR (CBIR) methods in conjunction with translation.

The main areas of study for a cross–language image retrieval assistant include:

- How well a system supports user query formulation for images with associated texts (e.g. captions or metadata) written in a language different from the native language of the users. This is also an opportunity to study how the images themselves could also be used as part of the query formulation process.

- How well a system supports query re–formulation, e.g. the support of positive and negative feedback to improve the user's search experience, and how this affects retrieval. This aims to address issues such as how visual and textual features can be combined for query reformulation/expansion.

- How well a system allows users to browse the image collection. This might include support for summarising results (e.g. grouping images by some pre-assigned categorization scheme or by visual feature such as shape, colour or texture). Browsing becomes particularly important in a CLIR system when query translation fails and returns irrelevant or no results.

- How well a system presents the retrieved results to the user to enable the selection of relevant images. This might include how the system presents the caption to the user (particularly if they are not familiar with the language of the text associated with the images, or some of the specific and colloquial language used in the captions) and investigate the relationship between the image and caption for retrieval purposes.

The interactive image retrieval task in 2004 concentrated on query re–formulation and this has been the focus of experiments in 2005 also, together with the presentation of search results. Groups were not set a specific retrieval goal to enable some degree of flexibility.

2.1 Experimental Procedure

Participants were required to compare two interactive cross–language image retrieval systems (one intended as a baseline) that differ in the facilities provided for interactive retrieval. For example, comparing the use of visual versus textual features in query formulation and refinement. As a cross-language image retrieval task, the initial query was required to be in a language different from the collection (i.e. not English) and translated into English for retrieval. Any text displayed to the user was also required to be translated into the user's source language. This might include captions, summaries, pre-defined image categories etc. ImageCLEF used a within–subject experimental design: users were required to test both interactive systems.

The same search task as 2004 was used: given an image (not including the caption) from the St Andrews collection of historic photographs, 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 (i.e. a high precision task or target search [2]).

The interactive ImageCLEF task is run similar to iCLEF 2003 using a similar experimental procedure. However, because of the type of evaluation (i.e. whether known items are found or not), the experimental procedure for iCLEF 2004 (Q&A) is also very relevant and we make use of both iCLEF procedures. The user–centered search task required groups to recruit a minimum of 8 users (native speakers in the source language) to complete 16 search tasks (8 per system). Images which users were required to find are shown in Fig. 1. Users are given a maximum of 5 mins only to find each image. Topics and systems were presented to the user in combinations following a latin–square design to ensure minimisation of user/topic and system/topic interactions.

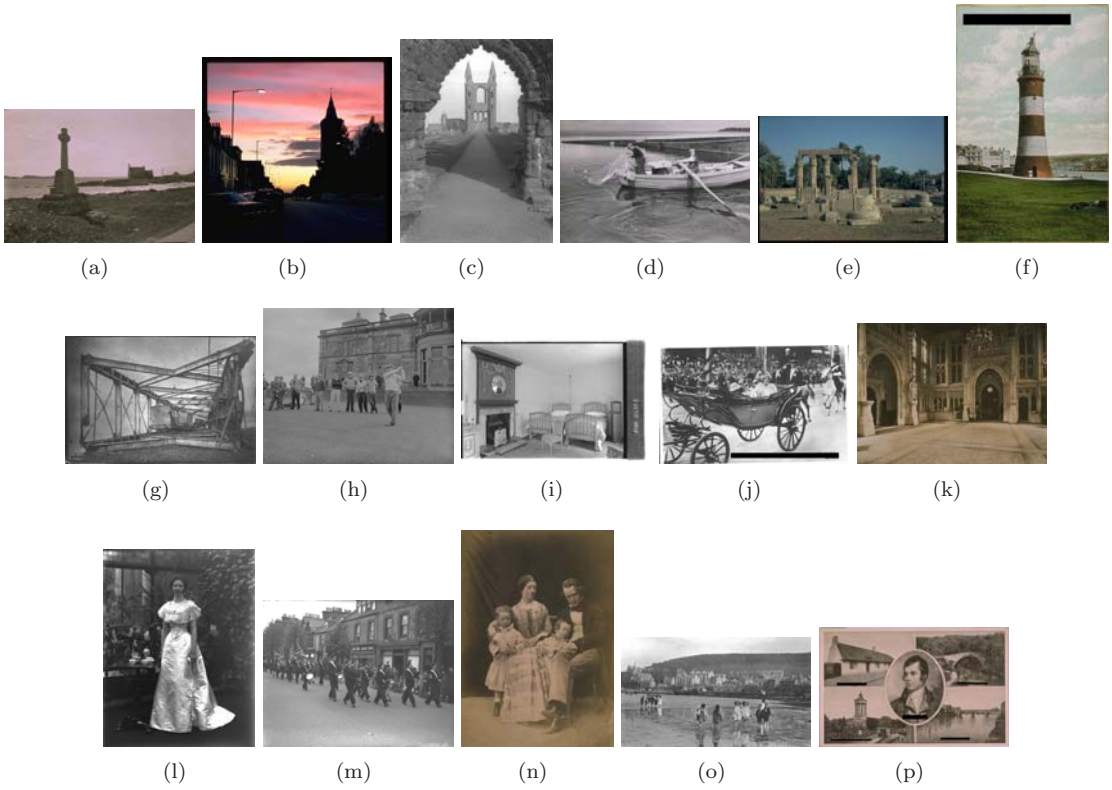

Figure 1: Example images given to participants for the user-centered retrieval task.

Participants were encouraged to make use of questionnaires to obtain feedback from the user

about their level of satisfaction with the system and how useful the interfaces were for retrieval. To measure the effectiveness and efficiency with which a cross–language image retrieval search could be performed, participants were asked to submit the following information: whether the user could find the intended image or not (mandatory), the time taken to find the image (mandatory), the number of steps/iterations required to reach the solution (e.g. the number of clicks or the number of queries - optional), and the number of images displayed to the user (optional).

2.2 Participating Groups

Although 11 groups signed up for the interactive task, only 2 groups submitted results: Miracle and the University of Sheffield. Miracle compared the same interface but using Spanish (European) versus English versions [8]. The focus of the experiment was whether it is better to use an AND operator to group terms of multi-word queries (in the English system) or combine terms using an OR operator (in the Spanish system). Their aim was to compare whether it is better to use English queries with terms conjuncted (which have to be precise and use the exact vocabulary - maybe difficult for a specialised domain like historical Scottish photographs) or to use the disjunction of terms 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). Their objective was to test the similarity of retrieval performance using both approaches.

Sheffield compared 2 interfaces with the same source language (Italian): one displaying search results as a list, the other organizing retrieved images into a hierarchy of text concepts displayed on the interface as an interactive menu [7]. The aim of the experiment was to determine the usefulness of grouping results using concept hierarchies and investigate translation issues in cross–language image search. Queries were translated using Babelfish and the entire user interface also translated to provide a working system in Italian.

2.3 Results and Discussion

Given only two submissions, conclusions that can be deduced from the interactive task are limited. However, the findings of individual groups were interesting and we summarise their main results to highlight the effectiveness of selected approaches. Miracle found results to be similar for both systems evaluated: English (69% of images found; 102 secs. average search time), Spanish (66% of images found; 113 secs. average search time). Based on investigation of the results and observation of users, a number of interesting points are made: that domain-specific terminology causes problems for cross–language searches (and therefore impacts far greater on queries with a conjunction of terms). In addition, translated Spanish query terms did not match caption terms also causing vocabulary mismatch. From questionnaires, users preferred the English version because the conjunction of terms often gave results users expected (i.e. a set of documents containing all query terms). Miracle also observed users extracting words from captions to further refine their search and user's commented on differences between the expected results of a search for a given keyword and those actually obtained. Users were also allowed to continue searching after the allotted time and in most cases found the relevant image in a short time (less than 1 minute).

The experiments undertaken by Sheffield also highlighted some interesting search strategies by users and problems with the concept hierarchies and interface for cross–language image retrieval. Quantitative results were similar using both a list of images and a menu generated from the concept hierarchies: list (53% of images found; 113 secs. average search time) and menu (47% images found; 139 secs. average search time). Overall users of the Sheffield systems found 82/128 relevant images and users of the Miracle system 86/128 images. The experiments undertaken by Sheffield observed negative effects on search, generation of the concept hierarchy and results display due to translation errors such as mis-translations and un-translated terms. Although based on effectiveness the menu appears to offer no difference compared to presenting results as a list, users preferred the menu (75% vs. 25% for the list) indicating this approach to be an engaging and interesting feature. In particular users liked the compact representation of search results offered by the menu compared to the ranked list.

3 Question Answering task

3.1 Experiment Design

Participating teams performed an experiment by constructing two conditions (identified as "reference" and "contrastive"), formulating a hypothesis that they wished to test, and using a common evaluation design to test that hypothesis. Human subjects were in groups of eight (i.e., experiments could be run with 8, 16, 24, or 32 subjects). Each subject conducted 16 *search sessions*. A search session is uniquely identified by three parameters: the human subject performing the search, the search condition tested by that subject (reference or contrastive), and the question to be answered. Each team used different subjects, but the questions, the assignment of questions to searcher-condition pairs, and the presentation order were common to all experiments. A latin-square matrix design was adopted to establish a set of presentation orders for each subject that would minimize the effect of user-specific, question-specific and order-related factors on the quantitative task effectiveness measures that were used. The remainder of this section explains the details of this experiment design.

3.1.1 Question set

Questions were selected from the **CLEF 2005 QA question set** in order to facilitate insightful comparisons between automatic and interactive experiments that were evaluated under similar conditions. The criteria to select questions was similar to those used in iCLEF 2004:

- **Answers should not be known in advance** by the human subjects; this restriction resulted in elimination of a large fraction of the initial question set.
- Given that the question set had to be necessarily small, we wanted to **avoid NIL questions** (i.e., questions with no answer. Ideally, it should be possible to find an answer to every question in any collection that a participating team might elect to search.
- We focused on **four question types** to avoid excessive sparseness in the question set: two question types that called for named entities as answers (PERSON and ORGANIZATION) and two question types that called for temporal or quantitative measures (TIME and MEASURE). The additional restriction of having answers in the largest number of languages forced us to include also some OTHER questions.

The final set of sixteen questions, plus four additional questions for user training, are shown in Table 1.

3.1.2 Latin-Square Design

One factor that makes reliable evaluation of interactive systems challenging is that once a user has searched for the answer to a question in one condition, the same question cannot be used with the other condition (formally, the learning effect would likely mask the system effect). We adopt a within-subjects study design, in which the condition seen for each user-topic pair is varies systematically in a balanced manner using a latin square, to accommodate this. This same approach has been used in the Text Retrieval Conference (TREC) interactive tracks [3] and in past iCLEF evaluations [6]. Table 2 shows the presentation order used for each experiment..

3.1.3 Evaluation Measures

In order to establish some degree of comparability, we chose to follow the design of the automatic CL-QA task in CLEF-2005 as closely as possible. Thus, we used the same assessment rules, the same assessors and the same evaluation measures as the CLEF QA task:

#	QA#	type	Question
1	0052	MEAS	How old is Jacques Chirac?
2	0105	PERS	Which professor from Bonn received the Nobel Prize for Economics?
3	0131	ORG	Which bank donated the Nobel Prize for Economics?
4	0143	MEAS	How many victims of the massacres in Rwanda were there?
5	0263	ORG	Which institution initiated the European youth campaign against racism?
6	0267	ORG	Which Church ordained female priests in March 1994?
7	0299	OTHER	What was the nationality of most of the victims when the Estonia ferry sank?
8	0362	ORG	Which airline did the plane hijacked by the GIA belong to?
9	0385	OTHER	What disease name does the acronym BSE stand for?
10	0386	ORG	Which country organized "Operation Turquoise"?
11	0397	PERS	Who was the Norwegian Prime Minister when the referendum on Norway's possible accession to the EU was held?
12	0522	TIME	When do we estimate that the Big Bang happened?
13	0535	PERS	Who won the Miss Universe 1994 beauty contest?
14	0573	MEAS	How many countries have ratified the United Nations convention adopted in 1989?
15	0585	MEAS	How many states are members of the Council of Europe?
16	0891	TIME	When did Edward VIII abdicate?
17	0061	ORG	Name a university in Berlin. *(training)*
18	0070	OTHER	Name one of the seven wonders of the world. *(training)*
19	0327	PERS	Which Russian president attended the G7 meeting in Naples? *(training)*
20	0405	OTHER	What minister was Silvio Berlusconi prior to his resignation? *(training)*

Table 1: The iCLEF 2005 question set

user	**search order** (condition: $A\|B$, question: $1 \ldots 16$)															
1	A1	A4	A3	A2	A9	A12	A11	A10	B13	B16	B15	B14	B5	B8	B7	B6
2	B2	B3	B4	B1	B10	B11	B12	B9	A14	A15	A16	A13	A6	A7	A8	A5
3	B1	B4	B3	B2	B9	B12	B11	B10	A13	A16	A15	A14	A5	A8	A7	A6
4	A2	A3	A4	A1	A10	A11	A12	A9	B14	B15	B16	B13	B6	B7	B8	B5
5	A15	A14	A9	A12	A7	A6	A1	A4	B3	B2	B5	B8	B11	B10	B13	B16
6	B16	B13	B10	B11	B8	B5	B2	B3	A4	A1	A6	A7	A12	A9	A14	A15
7	B15	B14	B9	B12	B7	B6	B1	B4	A3	A2	A5	A8	A11	A10	A13	A16
8	A16	A13	A10	A11	A8	A5	A2	A3	B4	B1	B6	B7	B12	B9	B14	B15

Table 2: iCLEF 2005 Condition and Topic Presentation Order.

- Human subjects were asked to designate a supporting document for each answer (we eliminated the exceptions allowed last year, as for instance building an answer from the information in two documents, because in practice no user exploited these alternative possibilities).
- Users were allowed to record their answers in whatever language was appropriate to the study design in which they were participating. For example, users with no knowledge of the document language would generally be expected to record answers in the question language. Participating teams were asked to hand-translate answers into the document language after completion of the experiment in such cases in order to facilitate assessment.
- Answers were assessed by the same assessors that assessed the automatic CL-QA results for CLEF 2005. The same answer categories were used in iCLEF as in the automatic CL-QA track: *correct* (valid, supported answer), *unsupported* (valid but not supported by the designated document(s)), *non-exact* or *incorrect*. The CLEF CL-QA track guidelines at http://clef-qa.itc.it/2005/guidelines.html provide additional details on the definition of these categories.
- We reported the same official effectiveness measures as the CLEF-2005 CL-QA track. Strict accuracy (the fraction of correct answers) and lenient accuracy (the fraction of correct plus unsupported answers) were reported for each condition. Complete results were reported to each participating team by user, question and condition to allow more detailed analyses to be conducted locally.

3.1.4 Suggested User Session

We set a maximum search time of five minutes per question, but allowed our human subjects to move on to the next question after recording an answer and designating supporting document(s) even if the full five minutes had not expired. We established the following typical schedule for each 3-hour session:

Orientation	10 minutes
Initial questionnaire	5 minutes
Training on both systems	30 minutes
Break	10 minutes
Searching in the first condition (8 topics)	40-60 minutes
System questionnaire	5 minutes
Break	10 minutes
Searching in the second condition (8 topics)	40-60 minutes
System questionnaire	5 minutes
Final questionnaire	10 minutes

Half of the users saw condition A (the reference condition) first, the other half saw condition B first. Participating teams were permitted to alter this schedule as appropriate to their goals. For example, teams that chose to run each subject separately to permit close qualitative assessment by a trained observer might choose to substitute a semi-structured exit interview for the final questionnaire. Questionnaire design was not prescribed, but sample questionnaires were made available to participating teams on the iCLEF Web site (http://nlp.uned.es/iCLEF/).

3.2 Experiments

Three groups submitted results:

University of Alicante. This group investigated how much context is needed to recognize answers accurately with a low-medium knowledge of the document language [5]. Their baseline system shows whole passages (maximum context) to users, while the experimental system shows only a clause (minimum context). Both systems highlight query terms, synonyms of query terms and candidate answers to facilitate the task.

University of Salamanca. Their focus has been exploring the use of free on-line machine translation programs for query formulation and presentation of results [9]. Both systems compared permit entering the query either in the user language or in the target language; in the first case, machine translation is applied to the query before searching the collection. In the reference system, results are displayed without translation; the contrastive system permits translating passages. Users were classified as having "poor" or "good" foreign language skills in four experiments, Spanish to English and Spanish to French.

UNED. This team has compared searching full documents with searching single sentences [4]. Both systems highlight fragments of the appropriate answer type to help locating the answer. In addition, the contrastive system filters out sentences which do not contain expressions of the appropriate answer type.

Group	Users	Docs	Experiment Condition	Accuracy	
				Strict	Lenient
Alicante	ES	EN	full passages	.44	.45
Alicante	ES	EN	clauses	.34	.34
Salamanca	ES	EN	good lang. skills / no translation	.50	.53
Salamanca	ES	EN	good lang. skills / translation	.56	.56
Salamanca	ES	EN	poor lang. skills / no translation	.36	.42
Salamanca	ES	EN	poor lang. skills / translation	.39	.45
Salamanca	ES	FR	good lang. skills / no translation	.66	.67
Salamanca	ES	FR	good lang. skills / translation	.69	.73
Salamanca	ES	FR	poor lang. skills / no translation	.63	.70
Salamanca	ES	FR	poor lang. skills / translation	.61	.66
UNED	ES	EN	documents	.53	.53
UNED	ES	EN	sentences with answer type filter	.45	.45

Table 3: iCLEF 2005 Q&A results.

Table 3 shows the official results for each of the five experiments. Readers are referred to the papers submitted by the participating teams for analyses of results from specific experiments.

4 Future work

Although iCLEF experiments continue producing interesting research results, which may have a substantial impact on the way effective cross-language search assistants are built, participation in this track has remain low across the five years of existence of the track. Interactive studies, however, remain as a recognized necessity in most CLEF tracks.

In order to find an explanation for this apparent contradiction, a questionnaire was created to establish reasons for low participation in the interactive ImageCLEF task and sent to all ImageCLEF participants. Seven participants returned their questionnaires and, out of these, 6 stated (the 7th participated in interactive ImageCLEF) their reason for not participating was lack of time, 5 lack of local resources and 4 that interactive experiments involved too much set-up time. Interactive experiments consume resources which many groups do not have.

We can think of a number of measures to solve this problem:

- Lowering the cost of participation. One approach is to provide a common task in which all groups participate, or use a shared multilingual document collection which can be accessed via an API, e.g. Flickr, Yahoo! or Google. This is only a partial solution, because the highest cost comes from recruting, training and monitorizing users for the searching sessions. An alternative is devising an experiment design in which search interfaces are deployed in real working environments, and then study the search logs of real users with real needs. This is a less controlled environment which could, nevertheless, provide a wealth of information about why and how users search in a cross-language manner.
- Adding value to the experimental setting. For instance, if we could work with online multilingual collections which have large user communities, setting up cross-language search interfaces for them has the additional appeal of being able to provide demonstrations which turn into useful web services for a significant set of web users.

We are currently contemplating the possibility of using a large-scale, web-based image database, such as Flickr (www.flickr.com), for iCLEF experiments. The Flickr database contains over five million images freely accesible via web, daily updated by a large number of users and available for all web users. These images are annotated by the authors with freely chosen keywords in a naturally multilingual manner: most authors use keywords in their native language, some combine more than one language. In addition, photographs have titles, descriptions, colaborative annotations, and comments in many languages. Participating groups would have the opportunity of building search interfaces not only for testing/demo purposes, but also to offer a useful web service with many potential users.

Acknowledgments

We are indebted to Richard Sutcliffe and Christelle Ayache for taking care of the Q&A assessments; Victor Peinado and Fernando López for their assistance with submission processing, Javier Artiles for maintaining the iCLEF web page, and Jianqiang Wang for creating the Systran translations that were made available to the iCLEF teams. Thanks also to Daniela Petrelli for the fruitful discussions about the iCLEF experimental design, and to Carol Peters and Doug Oard for their support. This work has been partially supported by the Spanish Government, project R2D2-Syembra (TIC2003-07158-C04-02).

References

[1] . P. Clough, H. Müeller, T. Desealers, M. Grubinger, t. Lehmann, J. Jensen and W. Hersh. The CLEF 2005 Cross-Language Image Retrieval Track, in this volume.

[2] . J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. Proceedings of the 13th International Conference on Pattern Recognition, 3:361–369, 1996.

[3] William Hersh and Paul Over. TREC-9 interactive track report. In *The Ninth Text Retrieval Conference (TREC-9)*, November 2000. http://trec.nist.gov.

[4] V. Peinado, F. López-Ostenero, J. Gonzalo and F. Verdejo. UNED at iCLEF 2005: automatic highlighting of potential answers In this volume.

[5] B. Navarro, L. Moreno-Monteagudo, E. Noguera, S. Vázquez, F. LLopis and A. Montoyo. "How much context do you need?" An experiment about the context size in Interactive Cross-Language Question Answering. In this volume.

[6] Douglas W. Oard and Julio Gonzalo. The CLEF 2003 interactive track. In Carol Peters, editor, *Proceedings of the Fourth Cross-Language Evaluation Forum*. 2003.

[7] Petrelli, D. and Clough, P.D. Concept Hierarchy across Languages in Text-Based Image Retrieval: A User Evaluation, In this volume.

[8] Villena-Román, R., Crespo-García, R.M., and González-Cristóbal, J.C. Boolean Operators in Interactive Search, in this volume.

[9] A. Zazo, C. Figuerola, J. Alonso and V. Fernández. iCLEF 2005 at REINA-USAL: Use of Free On-line Machine Translation Programs for Interactive Cross-Language Question Answering. In this volume.