iCLEF 2004 Track Overview: Interactive Cross-Language Question Answering

Julio Gonzalo* and Douglas W. Oard†

Abstract

For the 2004 Cross-Language Evaluation Forum (CLEF) interactive track (iCLEF), five participating teams used a common evaluation design to assess the ability of interactive systems of their own design to support the task of finding specific answers to narrowly focused questions in a collection of documents written in a language different from the language in which the questions were expressed. This task is an interactive counterpart to the fully automatic cross-language question answering task at CLEF) 2003 and 2004. This paper describes the iCLEF 2004 evaluation design, outlines the experiments conducted by the participating teams, and presents some initial results from analysis of official evaluation measures that were reported to each participating team.

1 Introduction

The design of systems to support information access depends on three fundamental factors: (1) the user's task, (2) the way in which the system will be used to achieve that task, and (3) the nature of the information being searched. In the Cross-Language Evaluation Forum, it is assumed that the information being searched is expressed in a different natural language (e.g., Spanish) than that chosen by the user to express their information needs to the system (e.g. English). In the CLEF interactive track (iCLEF), it is further assumed that the user will engage in an iterative search process using a system that is designed to support human-system interaction. In 2001, 2003, and 2003, iCLEF modeled the user's task as finding documents that were topically relevant to a written statement of the information need. In 2004 iCLEF adopted a new task; to find specific answers to narrowly focused questions.

The iCLEF evaluations have two fundamental goals: (1) to explore evaluation design, and (2) to permit contrastive evaluation of alternative system designs. These goals are somewhat in tension; the first inspires us to try new tasks, while the second would benefit from stability and continuity in the task design. Over the first three years of iCLEF, our focus was on progressive refinement of the evaluation design for a consistent task (finding topically relevant documents), and substantial progress resulted. Individual teams can continue to use the evaluation design that were developed at iCLEF over those three years, and evaluation resources that were produced over that period (e.g., official and interactive topical relevance judgments) can be of continuing value to both CLEF participants and to teams that subsequently begin to work on cross-language information retrieval.

When selecting a new task for iCLEF this year, we considered two options: (1) cross-language question answering, and (2) cross-language image retrieval. Ultimately, we selected cross-language question answering because there was a broader base of prior work on the evaluation of fully automated question answering systems to which we could compare our results. The Image CLEF

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

†Human-Computer Interaction Laboratory, College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD, 20742, USA, oard@glue.umd.edu

track did, however, also explore the design of an interactive image retrieval task this year. We therefore achieved the best of both worlds, with the opportunity to learn about evaluation design for both tasks. Readers interested in interactive image retrieval should consult the Image CLEF overview paper in this volume. In this paper, we focus on interactive Cross-Language Question Answering (CL-QA). The next section describes the iCLEF 2004 CL-QA experiment design. That is followed by sections describing the experiments and providing an overview of the results obtained by the participating teams. The paper concludes with some thoughts about future directions for iCLEF.

2 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.

2.1 Question set

Question selection proved to be challenging. We adopted the following guidelines to guide our choice of questions:

- We selected only questions from the **CLEF 2004 QA question set** in order to facilitate insightful comparisons between automatic and interactive experiments that were evaluated under similar conditions.
- The largest **number of questions** that could be accommodated in three hours were needed in order to maximize the reliability of the quantitative measures of task effectiveness. Our experience in previous years suggests that three hours is about the longest we can expect subjects to participate in a single day, and extending an experiment across multiple days would adversely affect the practicality of recruiting an adequate number of subjects. We chose to allow up to five minutes for each search. Once training time was accounted for, this left time for 16 questions during the experiment itself.
- **Answers should not be known in advance** by the human subjects. This restriction proved to be particularly challenging in view regardless of the breadth of cultural backgrounds that we expected among the participating teams in this international evaluation, resulting in elimination of a large fraction of the CLEF 2004 QA set (e.g., "What is the frequency unit?," "Who is Simon Peres?" and "What are Japanese suicide pilots called?,"). Two types of questions were found to be more often compatible with this restriction: temporal questions (e.g., "When was the Convention on the Rights of the Child adopted?") and measure questions (e.g., "How many illiterates are there in the world?" or "How much does the world population increase each year?").
- 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. Ultimately, we found that we had to limit this restriction to presence in both the Spanish and English

collections in order to get a sufficiently large number of questions from which to choose. Together, these cover four of the five experiments that were run (the fifth used the French collection).

- A small set of question cannot have a representative number of questions for each question type. To avoid averaging over tiny sets of different types of questions, we decided to focus on **four question types**. The CLEF QA set includes eight question types: LOCATION (e.g., "In what city is St Peter's Cathedral?"), MANNER (e.g., "How did Jimi Hendrix die?"), MEASURE (e.g., "How much does the world population increase each year?"), OBJECT (e.g., "What is the Antarctic continent covered with?"), ORGANIZATION (e.g., "What is the Mossad?"), PERSON (e.g., "Who is Michael Jackson married to?"), TIME (e.g., "When was the Cyrillic alphabet introduced?"), and OTHER (e.g., "What is a basic ingredient of Japanese cuisine?"). We selected 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) and sought to balance those four types of questions in the final set. Some iCLEF 2004 question types call for definitions rather than succinct facts (e.g., "What is the INCB?"). We decided to omit definition questions because we felt that evaluation might be difficult in an interactive setting (e.g., a user might combine information found in documents with their own background knowledge and then create answers in their own writing style that could not be judged using the same criteria as automatic QA systems).

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

#	QA#	type	Question
1	001	TIME	What year was Thomas Mann awarded the Nobel Prize?
2	109	MEAS	How many human genes are there?
3	314	PERS	Who is the German Minister for Economic Affairs?
4	514	ORG	Who committed the terrorist attack in the Tokyo underground?
5	122	MEAS	How much did the Channel Tunnel cost?
6	113	TIME	When did Latvia gain independence?
7	217	MEAS	How many people were declared missing in the Philippines after the typhoon "Angela"?
8	242	PERS	Who is the managing director of the International Monetary Fund?
9	511	TIME	When did Lenin die?
10	219	MEAS	How many people died of asphyxia in the Baku underground?
11	534	PERS	Who is the president of Burundi?
12	543	ORG	What is Charles Millon's political party?
13	646	ORG	Of what team is Bobby Robson coach?
14	506	TIME	When did the attack at the Saint-Michel underground station in Paris occur?
15	318	MEAS	How many people live in Bombay?
16	287	PERS	Who won the Nobel Prize for Literature in 1994?
17	002	PERS	Who is the managing director of FIAT? *(training)*
18	195	MEAS	How many pandas are there in the wild in China? *(training)*
19	505	PERS	Who is the Russian Minister of Finance? *(training)*
20	512	TIME	When did the Iranian Islamic revolution take place? *(training)*

Table 1: The iCLEF 2004 question set

2.2 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 [1] and in past iCLEF evaluations [2]. Table 2 shows the presentation order used for each experiment..

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 2004 Condition and Topic Presentation Order.

2.3 Evaluation Measures

In order to establish some degree of comparability, we chose to follow the design of the automatic CL-QA task in CLEF-2004 as closely as possible. Thus, we used the same assessment rules, the same assessors and the same evaluation measures as the CLEF QA task:

- Human subjects were asked to designate a supporting document for each answer. Automatic CL-QA systems were required to designate exactly one such document, but for iCLEF we also allowed the designation of zero or two supporting documents:
  - We anticipated the possibility that people might construct an answer from information found in more than one document. Users were therefore allowed to mark either one or two supporting documents for an answer. When two documents were designated, assessors were instructed to determine whether both documents together supported the answer.
  - Upon expiration of the search time, users might wish to record an answer even though time would no longer be available to identify a supporting document. In such cases, we allowed users to write an answer with no supporting document. Assessors were instructed to judge such an answer to be correct if and only if that answer had been found by some automatic CLEF CL-QA system.

  Users were not encouraged to use either option, and in practice there were very few cases in which they were used.

- 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 2004. 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/2004/guidelines.html provide additional details on the definition of these categories. Assessment in CLEF is distributed geographically on the basis of the document language, so some variation in the degree of strictness of the assessment across languages is natural. For iCLEF 2004, assessors reported that they sometimes held machines to a higher standard than they applied in the case of fully automated systems. For example, "July 25" was accepted as an answer to "When did the attack at the Saint-Michel underground station in Paris occur?" for fully automatic systems (because the year was not stated in the supporting document), but it was scored as inexact for iCLEF because the assessor believed that the user should have been able to infer the correct year from the date of the article.

- We reported the same official effectiveness measures as the CLEF-2004 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.

2.4 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 Experiments

Five groups submitted results: The Swedish Institute of Computer Science (SICS) from Sweden, the University of Alicante, the University of Salamanca and UNED from (Spain), and the University of Maryland from the USA. Four of the five groups had previously participated in iCLEF (the University of Salamanca joined the track this year). Somewhat surprisingly, all of the participants used interactive CLIR systems of fairly conventional designs; none adapted existing QA systems to support this task. In the remainder of this section, we briefly describe the experiment run at each site.

Alicante. The experiment compared two passage retrieval systems. In both systems, the query was formulated in Spanish, automatically translated into English before passage retrieval, and then passages were shown to the users in English (untranslated). The reference system also showed ontological concepts for the query and the passage, ranking passages with the same concepts as the query higher. The contrastive system showed syntactic-semantic patterns (SSP) for the query and for each verb in the passage. The hypothesis being tested was that for users with low English skills, it would be more useful to find the answer through SSPs than through the whole passage.

Maryland. Two types of summaries were compared. The first was an indicative summary consisting of three sentence snippets sampled from the beginning, the middle, and the end of a document that each contain at least one query term. That type of summary aims to provide users with a concise overview of the document in order to permit rapid judgments of relevance. The second was an informative summary with one longer passage automatically selected by the system. Both systems used variants of the UMD MIRACLE interactive CLIR system, and the hypothesis being tested was that informative summaries would be

more useful that indicative summaries for this task. Maryland was also interested in studying search behavior (query formulation, query refinement, user-assisted query translation, relevance judgment, and stopping criteria) for interactive CL-QA . The experiment involved eight native English speakers searching Spanish documents to answer questions written in English.

UNED. The UNED hypothesis was that a passage retrieval system that filtered out paragraphs that did not contain expressions of and appropriate type (named entities, dates or quantities, depending on the question) could outperform a baseline consisting of a standard information retrieval system (Inquery) that indexed and displayed Systran translations of the documents (i.e. performing monolingual searches over the translated collection). A second research goal was to establish a strong baseline for interactive CL-QA to be compared with automatic CL-QA in the context of CLEF.

Salamanca The Salamanca team experimented with a passage retrieval system in which machine translation was used to translate the query. They tested whether the possibility of on-demand access to a full documents would be more useful for CL-QA than display of a passage alone. Both systems included suggestion of query expansion terms; another goal of the experiment was to determine whether users would take advantage of that possibility in a question answering task.

SICS. SICS explored the effect of interactive query expansion using paired users (working on different questions) that could communicate within the pair (e.g., to discuss system operation or vocabulary selection). Additional research goals were to explore the nature of communication within pairs and the the effect of a "bookmark" capability on user confidence in the reported result. The SICS experiment was monolingual, with French questions and French documents; the human subjects were all native speakers of Swedish with moderate skills in French.

4 Results and Discussion

In this section, we present the official results, draw comparisons with comparable results from the CLEF-2004 CL-QA track, and describe some issues that arose with the assessment of submitted answers.

4.1 Official results

Table 3 shows the official results for each of the five experiments. The following points stand out from our initial inspection of these results:

- Three of the five experiments yielded differences in strict accuracy of approximately 1 answer out of 16 (0.0625% absolute), suggesting that the magnitude of detectable differences with this experiment design is likely appropriate for the types of hypotheses being tested. Conformation of this result must, however, await the results of statistical significance tests (e.g., analysis of variance) at each site.

- Five of the ten tested conditions yielded strict accuracy above 0.50, indicating that the interactive CL-QA task is certainly feasible. There may still be room for improvement, however; even in the best condition, more than 30% of the answers were either incorrect, inexact, or unsupported. Inter-assessor agreement studies would be needed, however, before we can quantify the magnitude of the further improvement that could be reliably measured with this experiment design.

- Remarkably, the system used in the condition that yielded the highest strict accuracy (0.69) was one of the simplest baselines: a standard document retrieval system performing mono-

Group	Users	Docs	Experiment Condition	Accuracy	
				Strict	Lenient
Maryland	EN	ES	indicative summaries	0.61	0.66
Maryland	EN	ES	informative summaries	**0.63**	0.66
UNED	ES	EN	doc. retrieval + Systran	**0.69**	**0.73**
UNED	ES	EN	passage ret. + entity filter	0.66	0.72
SICS	FR	FR	baseline	**0.27**	**0.41**
SICS	FR	FR	contrastive	0.19	0.28
Alicante	ES	EN	ontological concepts	0.38	0.50
Alicante	ES	EN	syntactic/semantic patterns	**0.45**	**0.56**
Salamanca	ES	EN	only passages	0.49	0.55
Salamanca	ES	EN	passages + full documents	**0.55**	**0.70**

Table 3: Official iCLEF 2004 results (bold: higher scoring condition).

lingual searches over machine translation results. This suggests that when user interaction is possible, relatively simple systems designs may suffice for CL-QA tasks.

- No evidence is yet available regarding the utility of more sophisticated question answering techniques (e.g., question reformulation or finding candidate answers in side collections) for interactive CL-QA because all iCLEF 2004 experiments employed fairly standard cross-language information retrieval techniques.

Readers are referred to the papers submitted by the participating teams for analyses of results from specific experiments.

4.2 Comparison with CLEF QA results

English was the only document language for which multiple iCLEF experiment results were submitted, so we have chosen to focus our comparison with the CLEF 2004 CL-QA track on cases in which English documents were used. Results from 13 automatic systems were submitted to the CLEF-2004 CL-QA track for English documents. We compared the results of the six iCLEF 2004 conditions in which English documents were used (two conditions from each of three experiments) with the results from those 13 automatic runs.

Participating teams in the CLEF-2004 CL-QA track automatically found answers to 200 questions, of which 14 were common to iCLEF. Table 4 compares the results of the automatic systems on these 14 questions with the results of the interactive conditions on all 16 topics.[1]

Most of the interactive conditions yielded strict accuracy results that were markedly better than the fully automatic systems on these questions. These large differences cannot be explained by the omission of two questions in the case of the automatic systems; correct answers to those two questions would increase the strict accuracy of the best automatic system from 0.36 to 0.44, which is nowhere near the strict accuracy of 0.69 achieved by the best interactive condition. Nor could language differences alone be used to explain the large observed differences between the best interactive and automatic systems since the same trend is present over the five question languages that were tried with the automatic systems.

[1]Removal of two topics from the interactive results would unbalance some conditions, so the interactive results include the effect of two questions that were not assessed for the automatic systems.

Group	question	docs	Run	Accuracy	
				strict	lenient
Automatic Systems (14 questions)					
IRST	IT	EN	irst042iten	0.36	0.36
IRST	IT	EN	irst041iten	0.29	0.29
DFKI	DE	EN	dfki041deen	0.14	0.21
BGAS	BG	EN	bgas041bgen	0.07	0.07
LIRE	FR	EN	lire042fren	0.07	0.07
DLTG	FR	EN	dltg041fren	0.07	0.07
EDIN	DE	EN	edin041deen	0.07	0.07
EDIN	FR	EN	edin042fren	0.07	0.07
LIRE	FR	EN	lire041fren	0	0
DLTG	FR	EN	dltg042fren	0	0
EDIN	DE	EN	edin042deen	0	0
EDIN	FR	EN	edin041fren	0	0
HELS	FI	EN	hels041fien	0	0
Average				**0.09**	**0.10**
Interactive Experiments (16 questions)					
UNED	ES	EN	doc retr + Systran	0.69	0.73
UNED	ES	EN	passage retr + entity filter	0.66	0.72
Salamanca	ES	EN	passage ret. + access full docs.	0.55	0.70
Salamanca	ES	EN	passage ret. - access full docs.	0.49	0.55
Alicante	ES	EN	syntactic/semantic patterns	0.45	0.56
Alicante	ES	EN	ontological concepts	0.38	0.50
Average				**0.53**	**0.62**

Table 4: Automatic vs. interactive experiments X → EN

This observed difference is particularly striking in view of our expectation that the question types that we chose for the interactive evaluation would be particularly well suited to the application automated techniques because the answers could be found literally in most cases. It seems reasonable to expect that the gap would be proportionally larger for more questions types that required a greater degree of inference.

It is also notable that human subjects received a larger relative benefit from lenient rather than strict scoring. The automatic results in Table 4 cannot accurately reveal differences smaller than 1 answer out of 14 (0.07). But half of the six interactive experiments exhibited differences at least that large, while only one of the eight (non-zero) automatic systems showed such a difference. We interpret this as an indication that lenient accuracy reflects characteristics of an answer than may be more prevalent in human question answering than in automatic question answering.

4.3 The Assessment Process

Richard Sutcliffe and Alessandro Vallin, who coordinated the iCLEF assessment process for English, offered the following observations about the process:

- Users made more elaborate inferences than machines. For example:

  `Q: When did Latvia gain independence? answer: 1991`

was judged correct even though the document said "*(..)breakup of the Soviet Union (..) in 1991*". In this case, the user inferred that Latvia was part of the Soviet Union. Another example of this effect is:

`Q: When did Lenin die? answer: January 20 1924`

The document states that "Friday is the 70th anniversary of Lenin's death." As it is dated on Saturday, 22 January 1994, the user could could have inferred the date. Of course, the user might also make mistakes that a machine would not; in this case, the date calculated by the user was off by one day, leading the answer to be scored as wrong.

- Sometimes inexact answers were provided when a more complete answer could be inferred. For example,

  `Q: When did the attack at the Saint-Michel underground station in Paris occur? answer: July 25`

  In this case, the user gave an incomplete answer "July 25," but the date of the document could have been used to accurately infer the year in which the event occurred. This could reflect a system limitation (the date of the document may not have been displayed to the user), or it may simply reflect a misunderstanding of the desired degree of completeness in the answer.

- The option to designate more than one supporting document was used only 9 times out of the 384 answers provided in the three experiments for which EN was the target language. In none of those 9 cases was it used correctly (i.e., no inference using combined information from both documents was appropriate). This suggests that this option may add an unhelpful degree of complexity to the evaluation process.

- People were more creative than machines regarding what constitutes a valid answer. For example, they might select "hundreds" as an answer, while automatic systems may fail to recognize such an imprecise expression as a possible answer.

- Manual translation of the answers into the document language after completion of the experiment introduced errors in a few cases. For example, a Spanish user correctly answered "15 mil millones de dolares," but it was translated with a typo "$15 billions" and therefore judged as inexact. When detected, these mistakes were corrected prior to generation of the official results (since it was not our objective to assess the manual answer translation process).

5 Conclusion and Future Plans

The iCLEF 2004 evaluation contributed a new evaluation design and results from five experiments in three language pairs with a total of 640 search sessions. The only similar evaluation of interactive question answering that we are aware of was the TREC-9 interactive track [1]. The iCLEF 2004 evaluation differs from the TREC-9 interactive track in two key ways: (1) iCLEF 2004 is focused on a cross-language task, while the TREC-9 interactive track focused on a monolingual task; and (2) iCLEF 2004 used questions and measures that facilitate comparison with an evaluation of automatic QA systems while the TREC-9 interactive track used more complex question types and document-oriented evaluation measures.

The iCLEF 2004 evaluations have already made a number of specific contributions, including:

- Developing a methodology to study user-inclusive aspects of CL-QA,

- Demonstrating that the accuracy of automatic QA systems is presently far below the accuracy that a typical user can obtain using a cross-language information retrieval system of fairly conventional design, and

- Establishing an initial baseline for the interactive CL-QA task, with a median across 8 tested conditions of about 50% strict accuracy for five-minute searches.

Much remain to be done, of course. Further analysis will be required before we are able to apportion the judged errors between the search and translation technologies embedded in the present systems. Moreover, we are now operating in a region where inter-assessor agreement studies will soon be needed if we are to avoid pursuing putative improvements that extend beyond our ability to measure their effect. Finally, there is a large design space that remains to be explored; no participating team has yet tried advanced techniques of the type normally used in fully automatic CL-QA systems in interactive systems.

Perhaps the most important legacy of iCLEF 2004 will be the discussions that it sparks about new directions for information retrieval research. How can we craft an evaluation venue that will attract participants with interests in both interactive and automatic CL-QA? What can we learn from the CLEF-2004 CL-QA evaluation that would help us design interactive CL-QA evaluations that reflect real application scenarios with greater fidelity? Given the accuracy achieved by interactive systems with a limited investment of the user's time, what applications do we see for fully automated CL-QA systems? With iCLEF 2004, we have gained a glimpse of these questions about our future, and we're looking forward to discussing them when we meet in Bath this September!

6 Acknowledgments

The authors would like to thank Alessandro Vallin and Richard Sutcliffe for serving as our liaison to the CLEF 2004 CL-QA track, for their help with assessments, and for sharing with us their insights into the assessment process. We are also grateful to Christelle Ayache for help with the French assessments, to Victor Peinado for file processing, to Fernando Lopez and Javier Artiles for creating the iCLEF 2004 Web pages, and to Jianqiang Wang for creating the Systran translations that were made available to the iCLEF teams.

References

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

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