

The CLEF 2003 Interactive Track

Douglas W. Oard* and Julio Gonzalo†

Abstract

The CLEF 2003 Interactive Track (iCLEF) was the third year of a shared experiment design to compare strategies for cross-language search assistance. Two kinds of experiments were performed: a) experiments in *Cross-Language Document selection*, where the user task is to scan a ranked list of documents written in a foreign language, selecting those which seem relevant to a given query. The aim here is to compare different translation strategies for an “indicative” purpose; and b) *Full Cross-Language Search* experiments, where the user task is to maximize the number of relevant documents that can be found in a foreign-language collection with the help of an end-to-end cross-language search system. Participating teams might choose to focus on any aspects of the search task (e.g., query formulation, query translation and/or relevance feedback). This paper describes the shared experiment design and briefly summarizes the experiments run by the five teams that participated.

1 Introduction

A Cross-Language Information Retrieval (CLIR) system, as that term is typically used, takes a query in some natural language and finds documents written in one or more other languages. From a user’s perspective, that is only one component of a system to help a user search foreign-language collections and recognize relevant documents. We generally refer to this situated task as *Multilingual Information Access*. The Cross-Language Evaluation Forum interactive track (iCLEF) in 2003 was the third occasion on which a community of researchers have used a shared experiment designed to compare strategies providing interactive support for the Multilingual Information Access process. As was the case in 2002, iCLEF 2003 included two tasks from which participating teams could select:

- Experiments in Cross-Language Document selection, where the user task is asked to scan a ranked list of documents that are written in a foreign language using some form of automated translation assistance, selecting those which seem to them to be relevant to a given topic. The aim here is to compare the degree to which different translation strategies are able to support the document selection process.
- Full Cross-Language Search experiments, where the user is asked to find as many relevant documents as possible with the help of a complete interactive CLIR system.

Seven teams registered for the track, from which five submissions were received. In Section 2 we describe the shared experiment design in detail, and in Section 3 we enumerate the participants and describe the hypotheses that they sought to test. Section 4 briefly recaps the official results; the preliminary analysis of these results can be found in each team’s paper. Finally, in Section 5 we make some observations about this year’s track and briefly discuss the prospects for the future of iCLEF.

*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

†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

2 Experiment Design

The basic design for an iCLEF 2003 experiment consists of:

- Two systems to be compared, usually one of which is intended as a reference system;
- A set of searchers, in groups of 8;
- A set of 8 topic descriptions, written in a language in which the searchers are fluent;
- A document collection in a different language (usually one in which the searchers lack language skills);
- A standardized procedure to be followed in every search session;
- A presentation order (i.e., a list of user/topic/system combinations which defines every search session and their relative order); and
- A set of evaluation measures for every search session and for the overall experiment, to permit comparison between systems.

Compared to iCLEF 2002, the main changes are the increase in the number of topics seen by each searcher from four to eight, the increase in the minimum number of searchers from four to eight, and (in order to keep the experiment duration reasonable) the decrease in the time per search from 20 minutes to 10. These changes reflect lessons learned in 2003, where statistical significance testing offered only a limited degree of insight with a more limited number of topics and searchers. In the remainder of this section, we describe these aspects in detail.

2.1 Topics

Topics for iCLEF 2003 were selected from those used for evaluation of fully automated ranked retrieval systems in the CLEF 2002 evaluation campaign. The main reason that we selected a previous year's topics was that it offered better insight into the number of relevant documents per topic and language, something that could not be guaranteed in advance with fresh CLEF 2003 topics.

The criteria for topic selection were:

- Select only broad (i.e., multi-faceted) topics.
- Select topics that had at least a few relevant documents in every document language, according to CLEF 2002 assessments.
- Discard topics that are too easy (for instance, when the presence of a proper noun is always correlated with relevance) or too difficult (for instance, when judging relevance needs a previous assessment on the topic).

These are the English titles and descriptions of the selected topics (description fields were also available, but are not shown here for space reasons):

```
<top>
<num> C100 </num>
<iCLEF> 1 </iCLEF>
<EN-title> The Ames espionage case </EN-title>
<EN-desc> Find documents that show the impact of the Ames espionage case
on U.S.-Russian relations. </EN-desc>
</top>

<top>
```

<num> C106 </num>
<iCLEF> 2 </iCLEF>
<EN-title> European car industry </EN-title>
<EN-desc> Find documents which report about the situation in the European car industry regarding the fall in sales (sales crisis) and possible countermeasures. </EN-desc>
</top>

<top>
<num> C109 </num>
<iCLEF> 3 </iCLEF>
<EN-title> Computer Security </EN-title>
<EN-desc> What is the status of computer security in regard to networked access? </EN-desc>
</top>

<top>
<num> C111 </num>
<iCLEF> 4 </iCLEF>
<EN-title> Computer Animation </EN-title>
<EN-desc> Find discussions of the impact of computer animation on the film industry. </EN-desc>
</top>

<top>
<num> C120 </num>
<iCLEF> 5 </iCLEF>
<EN-title> Edouard Balladur </EN-title>
<EN-desc> What is the importance for the European Union of the economic policies of Edouard Balladur? </EN-desc>
</top>

<top>
<num> C123 </num>
<iCLEF> 6 </iCLEF>
<EN-title> Marriage Jackson-Presley </EN-title>
<EN-desc> Find documents that report on the presumed marriage of Michael Jackson with Lisa Marie Presley or on their separation. </EN-desc>
</top>

<top>
<num> C133 </num>
<iCLEF> 7 </iCLEF>
<EN-title> German Armed Forces Out-of-area </EN-title>
<EN-desc> Find documents which report on political and juridical decisions on out-of-area uses of the Armed Forces of Germany. </EN-desc>
</top>

<top>
<num> C139 </num>
<iCLEF> 8 </iCLEF>
<EN-title> EU fishing quotas </EN-title>

<EN-desc> Find information about fishing quotas in the EU. </EN-desc>
</top>

We did not impose any restriction on the topic language; participating teams could pick any topic language provided by CLEF, or could prepare their own manual translations into any additional language that would be appropriate for their searcher population.

2.2 Document Collection

We allowed participants to search any CLEF document collection (Dutch, English, French, German, Italian, Spanish, Finnish or Swedish). To facilitate cross-site comparisons, we provided standard Machine Translations of the Spanish collection (into English) and of the English collection (into Spanish) for use by teams that found those language pairs convenient, in each case using Systran Professional 3.0.

2.3 Search Procedure

For teams that chose the full (end-to-end) search task, searchers were given a topic description written in a language that they could understand and asked to use one of the two systems to find as many relevant documents as possible in the foreign-language document collection. Searchers were instructed to favor precision rather than recall by asking them to envision a situation in which they might need to pay for a high-quality professional translation of the documents that they selected, but that they wished to avoid paying for translation of irrelevant documents.

The searchers were asked to answer some questions at specific points during their session:

- Before the experiment, about computer/searching experience and attitudes, and their language skills.
- After completing the search for each topic (one per topic).
- After completing the use of each system (one per system).
- After the experiment, comparing the two systems and soliciting and general feedback on the experiment design.

Every searcher performed eight searches, half with one system, and then half with the other. Each search was limited to 10 minutes. The overall time required for one session was approximately three hours, including initial training with both systems, eight 10-minute searches, all questionnaires, and two breaks (one following training, one between systems).

For teams that chose to focus solely on document selection, the experiment design was similar, but searchers were asked only to scan a frozen list of documents (returned by for some standard query by some automatic system) and select the ones that were relevant to the topic description from which the query had been generated.

2.4 Searcher/Topic/System Combinations

The presentation order for topics, searchers and systems was standardized to facilitate comparison between systems. We chose an order that was counterbalanced in a way that sought to minimize user/system and topic/system interactions when examining averages. We adopted a Latin square design similar to that used in previous iCLEF evaluations. The presentation order for topics was varied systematically, with participants that saw the same topic-system combination seeing those topics in a different order. An eight-participant presentation order matrix is shown in Table 1. Additional participants could be added in groups of 8, with the same matrix being reused as needed.

Searcher	Block 1	Block 2
1	System A: 1,4,3,2	System B: 5,8,7,6
2	System B: 2,3,4,1	System A: 6,7,8,5
3	System B: 1,4,3,2	System A: 5,8,7,6
4	System A: 2,3,4,1	System B: 6,7,8,5
5	System A: 7,6,1,4	System B: 3,2,5,8
6	System B: 8,5,2,3	System A: 4,1,6,7
7	System B: 7,6,1,4	System A: 3,2,5,8
8	System A: 8,5,2,3	System B: 4,1,6,7

Table 1: Presentation order for topics, and association of topics with systems.

2.5 Evaluation

In this section we describe the common evaluation measure used by all teams, and the data that was available to individual teams to support additional evaluation activities. These measures are identical to iCLEF 2002.

2.5.1 Data Collection

For every search (i.e., searcher/topic/system combination), two types of data were collected:

- The set of documents selected as relevant by the searcher. Optional attributes are the *duration* of the assessment process, the *confidence* in the assessment, and judgment values other than “relevant” (such as “somewhat relevant,” “not relevant,” or “viewed but not judged.”)
- The ranked lists of document identifiers created by the ranked retrieval system. One list was submitted by teams focusing on document selection; teams focused on query formulation and reformulation were asked to submit one ranked list for every query refinement iteration.

2.5.2 Official Evaluation Measure

The set of documents selected as relevant was used to produce the official iCLEF measure, an unbalanced version of van Rijsbergen’s F measure that we called F_α :

$$F_\alpha = \frac{1}{\alpha/P + (1 - \alpha)/R}$$

where P is precision and R is recall [2]. Values of α above 0.5 emphasize precision, values below 0.5 emphasize recall [1]. As in CLEF 2001, $\alpha = 0.8$ was chosen, modeling the case in which missing some relevant documents would be less objectionable than finding too many documents that, after perhaps paying for professional translations, turn out not to be relevant.

The comparison of average $F_{\alpha=0.8}$ measures across the two systems being tested provides a first order characterization of the effect of system differences on search effectiveness, but participating teams are encouraged to augment this comparison with additional measures based on the analysis of all available data (ranked lists for each iteration, assessment duration, assessment confidence, questionnaire responses, observational notes, statistical significance tests, etc.).

2.5.3 Relevance assessments

We provided relevance assessments by native speakers of the document languages for at least:

- All documents for which judgments were made by searchers (to support reliable computation of $F_{\alpha=0.8}$).

- The top 20 documents in every ranked list produced during a search iteration with an end-to-end search system.

All iCLEF 2003 relevance judgments were done by CLEF assessors immediately after assessing the CLEF 2002 pools. Only documents that had not been previously assessed in CLEF 2002 were specifically judged for iCLEF 2003.

3 Participants

Seven teams expressed interest in participating, and five teams submitted experiment results: University of Alicante (Spain), SICS (Sweden), University of Maryland (UMD, USA), a team formed jointly by BBN Technologies and the University of Maryland (BBN/UMD, USA) and UNED (Spain). Three groups focused on document selection strategies:

- **SICS** (Sweden). The SICS iCLEF experiments were, as last year, centered on trying to measure differences between assessing texts in one's native language and one in which the searcher has a near-native competence. The hypothesis being tested was whether intra-subject differences between native and near-native languages were significantly different; it seemed reasonable to expect that assessment would be slower and less reliable in a foreign language, even one in which the subject is fluent on a professional level. This year SICS used a system developed as part of the CLARITY project, components of which were also used in last year's iCLEF experiments. One of the salient features of the interface is a panel in which the user can store results from the assessment process. Some debriefing questions were added to last year's protocol to investigate the user's confidence in their judgments.
- **University of Alicante** (Spain) compared a query-oriented passage extraction system (presented at iCLEF 2002) with a new automatic extraction approach based on syntactic and semantic patterns based on the main verb of the sentence and its arguments. Thus, such patterns show only the basic information of each sentence. The language best known by the searchers was Spanish; the document language was English, a language in which the searchers self-reported passive language abilities (i.e, recognition, but not production). The goal of the experiment was to discern which of the approaches would best support rapid and accurate selection of relevant documents.
- **BBN Technologies/University of Maryland** (USA) compared the use of brief summaries constructed using automatic headline generation with the use of the first 40 words from each story as the summary. The document language was Spanish, for which the eight searchers self-reported little or no fluency. The searchers were fluent in English, so the standard Systran translations were used as a basis for both conditions. Headlines were automatically produced for each document by removing grammatical constituents from a parse tree of the lead sentence of a document until a length threshold was been met. The hypothesis being tested was that headlines could support more rapid assimilation of the topic of a document without adverse effects on accuracy.

The other two groups experimented with full cross-language searches:

- **University of Maryland** (USA). The focus of Maryland's full-system experiment was on query formulation and iterative query reformulation. The hypothesis was that providing greater insight into and control over the query reformulation process could improve the usability of a search system, yielding greater accuracy with less reformulation effort. Participants interacted with two systems. One provided descriptions for the available translations and allowed control over which translations were used, The other performed fully automatic query translation. The query language was English and the document language was Spanish; searchers self-reported little or no fluency in Spanish.

- **UNED** (Spain). The UNED experiment tested whether document summaries based on phrase translation (which UNED used in document selection experiments in 2001 and in query reformulation experiments in 2002) could also be used as the basis for a document translation approach to interactive cross-language searching. The phrase translation summaries contained only 30% as many words as the original documents, and could be generated two orders of magnitude faster than full machine translation. Users performed searches with two systems. In one, the system generated possibly relevant noun phrases in response to a query, and the user picked some of those phrases for automatic translation, and the system then used the translated phrases to search the documents. The nature of the user interaction in the second system was the same, but instead of translating the selected noun phrases, the search was performed on the phrase translation summaries in the query language. Spanish was the query language and English was the document language. The hypothesis was that searching phrase translation summaries (which were needed in any case for display) could yield comparable search effectiveness to the approach based on query translation.

4 Results and Discussion

The available official results for the $F_{\alpha=0.8}$ measure are shown in Table 2.¹

Group	Experiment Condition	$F_{\alpha=0.8}$
Experiments in Query formulation and refinement		
Maryland	automatic query translation	.20
Maryland	assisted query translation	.23
UNED	query translation	.29
UNED	(summarized) document translation	.29
Experiments in Document selection		
SICS	foreign language docs	
SICS	native language docs	
Alicante	passages	.45
Alicante	patterns	.44
BBN/UMD	First 40	.47
BBN/UMD	Hedge	.38

Table 2: Official iCLEF 2003 results.

5 Conclusions

The iCLEF design evolved rapidly over the first two years of the track; this year’s design included only evolutionary improvements over last year’s. Growth in the number of participating teams now seems to be leveling off; five teams participated in 2002 and in 2003, although participation by a sixth team would have been likely if we had been able to provide ranked lists for use in the document selection task a bit sooner. This, therefore seems like a good time to think about the future of iCLEF.

First, we should make the point that iCLEF is not, and never has been, the only venue for evaluation of interactive CLIR; several individual researchers have run well designed studies to explore one aspect or another of this topic. Rather, the unique strength of iCLEF is in the community that it draws together. Advancing the state of the art in interactive CLIR requires expertise in several areas, including information retrieval, computational linguistics, and human-computer interaction. Few research teams can draw on such a broad range of expertise, iCLEF

¹Formatting difficulties prevented official scoring of the SICS results before this paper was due. These results will be reported at the workshop and in the final track report in the post-workshop proceedings.

includes two or more teams with interests in each. Moreover, as with many specialized tracks, iCLEF serves to enrich the dialog at CLEF by bringing in researchers with new perspectives that might not otherwise participate. The evaluation framework that we have evolved is indeed a useful and important contribution, but we expect that the greatest legacy of iCLEF will result from the discussions we have had and the ideas we have shared.

Where next for iCLEF? One possibility is to continue the process we have started. Our assessment process leverages the work already being done for CLEF, and it has the felicitous side effect of contributing additional relevance judgments that may be useful to those who are interested in studying the assessment process. Research teams around the world are now working on interactive CLIR, and iCLEF provides a natural venue in which they can report their results and share their ideas. After iCLEF 2002 we discussed some related tasks that we might also try; searching narrow (single-aspect) topics and interactive cross-language question answering were two of the ideas we considered. Ultimately, we decided that the community's best interests would be served by a year of relative stability in the evaluation design, allowing the participating research teams to build on their results from last year. But there is no reason why we should not explore these ideas, and others, again.

The future of iCLEF is, of course, to some extent bound up with the future of CLEF itself. Here, there are two countervailing forces to consider. iCLEF adds a valuable dimension to CLEF, but it also competes for resources with other good ideas. In a world with constrained resources, choices will need to be made. A spirit of exploration has been one of the hallmarks of CLEF, and we should not be afraid to explore radical ideas that may take us in interesting new directions. If cross-language question answering yields interesting results this year, then perhaps we might try an interactive task within the question-answering track next year. If cross-language caption-based image retrieval works well, why not interactive caption-based searches for images? If cross-language spoken document retrieval goes in interesting directions, perhaps interactive searching for foreign-language speech would be the next natural challenge. Ultimately, each of these tracks seeks to meet the needs of real users, so it is natural to expect that each will want to involve users in their research at some point. The interactive CLIR community is still relatively small, so we can not hope to go in all of these directions at once. But as is said that a journey of a thousand li (a measure of distance in ancient China) begins with a single step. Over the past three years, we have taken that step, and as a group we have achieved some interesting and important results. Now is the time to think about the next step.

Acknowledgments

We are indebted to many people that helped along the organization of this iCLEF track: Fernando López wrote the evaluation scripts and maintained the web site and distribution list, Martin Braschler created the assessment pools; Ellen Voorhees, and Djuna Franzén coordinated relevance assessments, and Jianqiang Wang provided Systran translations for English and Spanish collections. Finally, we also want to thank Carol Peters for her permanent support and encouragement.

References

- [1] Douglas Oard. Evaluating cross-language information retrieval: Document selection. In Carol Peters, editor, *Cross-Language Information Retrieval and Evaluation: Proceedings of CLEF 2000*, Springer-Verlag Lecture Notes in Computer Science 2069, 2001.
- [2] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.