The CLEF 2002 Interactive Track.

Julio Gonzalo and Douglas W. Oard

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
WWW home page: http://sensei.lsi.uned.es/~julio/
and
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.edu,
WWW home page: http://www.glue.umd.edu/~oard/

Abstract. In the CLEF 2002 Interactive Track, research groups interested in the design of systems to support interactive Cross-Language Retrieval used a shared experiment design to explore aspects of that question. Participating teams each compared two systems, both supporting a full retrieval task where users had to select relevant documents given a (native language) topic and a (foreign language) document collection. The two systems being compared at each site should differ in (at least) one of these aspects: a) support for document selection (how the system describes the content of a document written in a foreign language), b) support for query translation (how the system interacts with the user in order to obtain an optimal translation of the query), and c) support for query refinement (how the system helps the user refine their query based on previous search results). This paper describes the shared experiment design and summarizes preliminary results from the five teams that submitted runs.

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*. To emphasize the importance of interactive mechanisms in a situated, user-centered cross-language search, we might refer to systems that support that task as *Cross-Language Search Assistants*.

The Cross-Language Evaluation Forum's (CLEF) interactive track (iCLEF) aims to develop shared experiment designs that will allow research teams to compare their strategies for cross-language search assistance. In the first year of the track, iCLEF 2001 focused on comparing document selection strategies (i.e., approaches to facilitate fast and accurate relevance judgments for documents that the user could not read without assistance). The experiment guidelines combined CLEF resources (including the CLEF documents and topic descriptions, ranked lists from CLEF 2000, and relevance assessments made by native speakers in every document language) with the within-subject quantitative user study design that was used for many years in the TREC interactive track. The success of that evaluation [3] led to a decision to continue the track in the CLEF 2002 evaluation campaign. This paper describes the design of the iCLEF 2002 evaluation and presents results on the relevance assessment process.

The CLEF 2002 interactive track retained provisions for studying document selection, but added a new focus on comparing interactive query formulation and reformulation. Unlike document selection, which can reasonably be studied as an isolated process, query formulation and reformulation only make sense as part of a larger process; queries have no inherent value beyond their effect on search results. We therefore chose an experiment design that supported comparison of complete interactive CLIR systems, while retaining provisions for more focused experiments. Each participating team compared two systems that differed in one or more aspects of the interaction with the user. Five groups submitted results: Swedish Institute of Computer Science (SICS, Sweden), University of Sheffield (UK), University of Alicante and University of Jaen (Spain), University of Maryland (USA) and Universidad Nacional de Educación a Distancia (UNED, Spain).

In Section 2 we describe the experiment design in detail, and in Section 3 we enumerate the participants and the hypotheses that they sought to test. In Section 4, we summarize the experiments run by each team and briefly discuss the suitability of our current experiment design. Finally, in Section 5 we draw some conclusions and describe the prospects for future iCLEF evaluations.

2 Experiment Design

The basic design for an iCLEF 2002 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 4;
- A set of 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 that 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.

In the remainder of this section, we describe these aspects in detail.

2.1 Topics

Topics for iCLEF 2002 were selected from those used for evaluation of fully automated ranked retrieval systems in the CLEF 2001 evaluation campaign. The main reason that we selected a previous year's topics was that it allowed more time between topic release and the submission deadline, an important factor when performing user studies.

The criteria for topic selection were:

- Select only broad (i.e., multi-faceted) topics. In iCLEF 2001, we observed that narrow (single-faceted) topics tended to have very few relevant documents, which made evaluation measures based on the fraction of relevant documents retrieved less insightful.
- Select topics that had at least 8 relevant documents in every document language, according to CLEF 2001 assessments.
- Select topics that could reasonably be expected to be found in collections from different years. This provided a degree of assurance that the new CLEF 2002 document collections could also be used by participating teams (the Finnish collection is mainly news from 1995, while the others are mainly 1994).

These restrictions were satisfied by eight topics, from which four were selected as iCLEF 2002 topics:

```
<num> C053 </num>
<EN-title> Genes and Diseases </EN-title>
<EN-desc> What genes have been identified that are the source of or
contribute to the cause of diseases or developmental disorders in
human beings? </EN-desc>
<EN-narr> A document that identifies a gene or reports that a gene
has been discovered that is the source of any type of disease,
syndrome, behavioral or developmental disorder in humans is
relevant. Any document that reports the discovery of a defective
gene that causes problems in humans is relevant, but reports of
diseases and disorders that are caused by the absence of a gene are
not relevant. </EN-narr>

<num> C056 </num>
<EN-title> European Campaigns against Racism </EN-title>
<EN-desc> Find documents that talk about campaigns against racism in
Europe. </EN-desc>
<EN-narr> Relevant documents describe informative or educational
campaigns against racism (ethnic or religious, or against
immigrants) in Europe. Documents should refer to organized campaigns
rather than reporting mere opinions against racism. </EN-narr>
```

```
<num> C065 </num>
<EN-title> Treasure Hunting </EN-title>
<EN-desc> Find documents about treasure hunters and treasure hunting
activities. </EN-desc>
<EN-narr> Identify types of current treasure hunting activities such
as searching for gold, digging for buried relics, or searching
underwater for sunken galleons. </EN-narr>

<num> C080 </num>
<EN-title> Hunger Strikes </EN-title>
<EN-desc> Documents will report any information relating to a hunger
strike attempted in order to attract attention to a cause. </EN-desc>
<EN-narr> Identify instances where a hunger strike has been
initiated, including the reason for the strike, and the outcome if
known. </EN-narr>
```

and one was selected as a training topic:

```
<num> C086 </num>
<EN-title> Renewable Power </EN-title>
<EN-desc> Find documents describing the use of or policies regarding
"green" power, i.e., power generated from renewable energy
sources. </EN-desc>
<EN-narr> Relevant documents discuss the use of renewable energy
sources such as solar, wind, biomass, hydro, and geothermal sources.
Low emission vehicles as for example electric or CNG cars are not
relevant. Fuel cells are not relevant unless their fuel qualifies as
renewable. </EN-narr>
```

The number of relevant documents for these topics in the CLEF 2001 pools can be seen in Table 1.

Topic	Dutch	English	French	German	Italian	Spanish
				(SDA + Spiegel)		
53	27	36	13	17	37	33
56	8	10	44	20	24	137
65	69	15	13	47	15	74
80	93	56	31	62	84	245
86	50	82	36	58	31	56

Table 1. Number of relevant documents for iCLEF 2002 topics in previous pools.

We did not impose any restriction on the topic language. Participants 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 and Swedish). To facilitate cross-site comparisons, we provided standard Machine Translations of the German 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. A fraction of the German collection (the Frankfurter Rundschau set) was discarded for iCLEF experiments because both Systran and an alternative system that we tried produced no output for a significant fraction of those documents. The other collections were used in their entirety.

2.3 Search Procedure

For teams that chose the end-to-end experiment design, 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, about system comparison and general feedback on the experiment design.

These questions were normally posed using questionnaires that closely followed the design of the questionnaires used in iCLEF 2001. This year, however, we did not require the use of standardized questionnaires; Participating teams could adapt the examples that we provided to their particular experiment conditions in whatever way they wished.

Every searcher performed four searches, first for two topics using one system and for the remaining two topics with the other system. Each search was limited to 20 minutes. The overall time required for one session was approximately three hours, including initial training with both systems, four 20-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. This is essentially the iCLEF 2001 task.

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 the TREC interactive track. 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 2.[1] The minimum number of participants was set at 4, in which case only the top half of the matrix would be used. Additional participants could be added in groups of 8, with the same matrix being reused as needed.

Searcher	Block 1	Block 2	Searcher	Block 1	Block 2
1	System 1: 1-4	System 2: 3-2	**5**	System 1: 4-2	System 2: 1-3
2	System 2: 2-3	System 1: 4-1	**6**	System 2: 3-1	System 1: 2-4
3	System 2: 1-4	System 1: 3-2	**7**	System 2: 4-2	System 1: 1-3
4	System 1: 2-3	System 2: 4-1	**8**	System 1: 3-1	System 2: 2-4

Table 2. 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.

Data Collection For every search (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 focusing on query formulation and reformulation were asked to submit one ranked list for every query refinement iteration.

[1] This table was prepared before the topics were chosen, and some participating teams refer to topics numbered 1–4 in their papers. The mapping for iCLEF 2002 is 1=C053, 2=C065, 3=C056, 4=C080.

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 [4]. Values of α above 0.5 emphasize precision, values below 0.5 emphasize recall [2]. 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. For the same reason, documents judged as "somewhat relevant" are treated as not relevant for computing $\alpha = 0.8$.

The comparison of average $F_{\alpha=0.8}$ measures for both systems provides the official, first order differentiation of systems. All complementary material (ranked lists for each iteration, assessment duration, assessment confidence, questionnaire responses, observational notes, etc.) can be used by participating groups as a basis for further analysis.

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

- All documents manually selected by searchers (to compute $F_{\alpha=0.8}$).
- The first 20 documents in all iterative rankings produced along every search process.

For the CLEF 2001 document languages (English, German, Italian, Spanish, Dutch, and French) we already had some assessments available from the CLEF 2001 pools. In the case of Finnish and Swedish, all assessments had to be done from scratch. All iCLEF 2002 relevance judgments were done by CLEF assessors immediately after assessing the CLEF 2002 pools.

3 Participants

Six teams expressed interest in participating, and five teams submitted experiment results: three that had participated in iCLEF 2001 (Sheffield, Maryland, and UNED), and two new teams (SICS and Alicante/Jaen). Both newcomers focused on the document selection subtask:

- **Alicante/Jaen** compared full machine translations (as the reference condition) with topic-oriented summaries of the same translations, containing the title and the most relevant paragraph for the topic being searched (as the contrastive condition). They used Spanish as the topic language, and English as the document language.

- **SICS** tested a hypotheses that assessing documents in one's native language would be less work than assessing documents in another language, even if that language is relatively well mastered. Therefore they used one topic language (Swedish) and two document languages: English and Swedish. Twelve Swedish users with high English skills participated in the experiment. The users were presented with prefabricated ranked lists of search results in an interface which allowed them to view each document and assess it for relevance. The ranked lists were either from the Swedish or the English CLEF collection, forming the two conditions being tested (native language versus foreign language assessments).

The other three groups focused on the query formulation and refinement aspects of interactive searches:

- **Maryland** used four searchers in their official submission to compare user-assisted query translation with a fully automatic approach. An additional eight searchers performed the same experiment with a smaller collection. The hypothesis being tested was that user-assisted query translation could improve search effectiveness. The document language was German, and the topic language was English. For the user-assisted query translation condition, searchers were provided two types of cues about the meaning of each translation: a list of other words sharing a common translation (potential synonyms) and a sentence in which the word was used in a translation-appropriate context selected from a word-aligned parallel corpus.
- **Sheffield** used four users with a prototype system being developed jointly by Sheffield, SICS, and the University of Tampere (Finland) to compare user-assisted translation with a fully automatic approach. The hypothesis being tested was that user-assisted query translation could improve search effectiveness. The search engine was created by Tampere using a modified version of the Inquery search system. The interface was designed by SICS and Sheffield based on interviews and observations of users with CLIR needs.
- **UNED** used eight searchers to compare a reference system using words as units for query formulation and refinement with a contrastive system using phrases. The hypothesis being tested was that phrases as interactive query formulation units could provide enough context information for accurate automatic translation, as an alternative to word-by-word user-assisted translation.

4 Results and Discussion

The official $F_{\alpha=0.8}$ measure for all systems is shown in Table 3[2]. A detailed discussion of each of the experiments can be found elsewhere in these proceedings. Most experiments showed substantial differences between the systems being

[2] The Sheffield results shown here are based on recomputation at Sheffield. Format problems in the submitted results precluded automatic official scoring.

compared, suggesting that there is a good deal to be learned from the detailed analysis reported in each team's paper.

Group	Experiment Condition	$F_{\alpha=0.8}$
Experiments in Query formulation and refinement		
Maryland	automatic query translation	0.34
Maryland	user-assisted query translation	0.50
Sheffield	automatic query translation	0.20
Sheffield	user-assisted query translation	0.26
UNED	word-based query translation	0.23
UNED	phrase-based query translation	0.37
Experiments in Document selection		
SICS	foreign language docs	0.36
SICS	native language docs	0.65
Alicante/Jaen	Systran full translations	0.22
Alicante/Jaen	Systran title + best passage	0.32

Table 3. Official iCLEF 2002 results.

German and English are the two languages for which a) there were available pools from CLEF 2001, and b) participants ran end-to-end interactive cross-language sessions to contribute new documents to the assessment pools (either because they were selected by the searchers as relevant or because they appeared near the top of some ranked lists during the search processes).

Voorhees has found that manual TREC runs (those which include any form of human intervention in the search process) often find documents that are not present in assessment pools generated from the output of automatic systems [5]. The CLEF 2001 pools (produced from 198 submitted runs) were already large and stable [1], so the iCLEF 2002 assessment pools provided us with an opportunity to explore this issue in a cross-language search context. We observed a similar effect. Table 4 summarizes the additional assessments and the additional relevant documents found with the new assessments.

In the case of the SDA and Der Spiegel subset of the German collection used in our evaluation, the large number of query reformulation iterations produced enormous pools, but only increased the set of known relevant documents by 10%. The newly judged pools were substantially smaller for English, but the set of known relevant documents still was increased by 12%. A plausible explanation is that, when a query formulation produces seemingly good results, searching time is primarily spent in the process of selecting documents from the ranked list returned by the system. When the query does not produce good results, time is spent in iterative query reformulations which enlarge the document pool. Hence, the harder the query, the larger the pool.

From this, we can conclude that although human searchers do find relevant documents that automatic systems miss, the search strategies that they employed resulted in many more non-relevant documents. This is a classic recall-precision tradeoff. It is important, of course, to caveat this observation by pointing out that we explored only a limited range of conditions (in particular, 20 minute searches for broad topics).

Another question that we might ask is whether number of documents added to the assessment pools is correlated with the number of relevant documents contained in those pools. As Table 4 indicates, there does seem to be a weak negative correlation; the topic with the fewest newly discovered relevant documents generated the largest assessment pools, for example. Our experiment design required relevance assessments for only 4 topics, so the assessment costs were not prohibitive in this case. But these observations may be helpful as we design future interactive CLIR studies.

Topic	German CLEF		iCLEF add-on	
	(SDA+Der Spiegel)		(SDA+Der Spiegel)	
	assessed	relevant	assessed	relevant
53	220	17	225	5 (+30%)
56	230	20	465	5 (+25%)
65	249	47	835	0 (=)
80	118	62	450	6 (+10%)

Topic	English CLEF		iCLEF add-on	
	assessed	relevant	assessed	relevant
53	456	36	22	1 (+3%)
56	626	10	419	1 (+10%)
65	613	15	233	10 (+67%)
80	578	56	250	2 (+4%)

Table 4. Contribution of interactive runs to CLEF 2001 pools.

5 Conclusions

Together, the five teams that participated in iCLEF 2002 had 38 searchers perform 158 searches in four document languages to test a broad range of hypothesis related to the design of cross-language search assistance systems. To the best of our knowledge, this is the largest multilingual information access user study ever performed. We therefore believe that the results obtained by the participating teams will be a rich source of evidence from which we can learn more about the way cross-language information retrieval technology will ultimately be used. Perhaps even more importantly, we have enriched our understanding of the design of user studies for end-to-end cross-language search assistance systems, and have expanded the community of researchers that share an interest in this important question.

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, Michael Kluck, Eija Airio and Jorun Kugelberg provided native relevance assessments; and Jianqiang Wang and Dina Demner-Fushman provided Systran translations for the German and English collections. Finally, we also want to thank Carol Peters for her continued support and encouragement.

References

1. M. Braschler. CLEF 2001 - overview of results. In *Evaluation of Cross-Language Information Retrieval Systems. Proceedings of CLEF 2001: revised papers*, Springer-Verlag LNCS 2406, 2002.
2. 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.
3. Douglas W. Oard and Julio Gonzalo. The CLEF 2001 interactive track. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Proceedings of CLEF 2001, Springer-Verlag LNCS Series 2069*, 2002.
4. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.
5. Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.