*i*CLEF 2004 – Guidelines

Track: The iCLEF challenge

Build a system that will allow real people to find information that is written in languages that they have not mastered. Then measure how well representative users are able to use that system.

The goal of **iCLEF** , thn, is to study the interactive aspects of Cross-Language Information Retrieval systems. Standard CLEF, NTCIR and Cross-Language TREC tasks evaluate the ability of systems to automatically retrieve target-language(s) documents from source-language queries; **iCLEF** evaluates how well systems **help users locate and identify** relevant foreign-language information.

iCLEF 2004 task: Interactive Cross-Language Q&A

This year, the interactive CLEF track will study the problem of Cross-Language Question Answering (CL-QA) from a user-inclusive perspective. The challenge is twofold:

- From the point of view of Cross-Language QA as a user task, **how well systems help users locate and identify answers to a question in a foreign-language document collection?**
- From the point of view of QA as a machine task, **how well interaction with the user helps a Cross-Language QA system retrieve better answers?**

For a CL-QA system, the issue is how best can the QA system interact with the user to obtain details about a question that facilitate the automatic search for an answer in the document collection. For instance, in case of ambiguity, the system may request additional information from the user, avoiding incorrect translations (for translation ambiguity) or incorrect inferences (for semantic ambiguity).

For a Cross-Language search system, the issue is how a system can best assist a user with the task of finding and recognizing the answer to a question by searching the document collection. In monolingual searches, users can often easily find answers using standard document or passage retrieval systems. The cross-language case seems, however, to demand much more assistance from the system:

- How can interaction optimize the translation of the query?
- How can the system best show the contents of foreign-language documents for answer location tasks?
- How can the system help query refinement for a CL QA task?

We welcome research teams with interests in cross-language information retrieval, human-computer interaction, question answering, and machine translation. The organizers hope to foster synergies between interested parties, so expertise in one or more of these fields should be sufficient to participate.

Research teams participating in iCLEF are supposed to study some of the issues above by **comparing two systems in a CL QA search task** involving a number of topics (provided by iCLEF) and a number of searchers (recruited locally by the participant team). The two systems should differ in the facilities provided for any of the tasks listed above. The iCLEF experiment design will allow groups to estimate the effect of system differences by suppressing the (additive) effects of participant and topic, and by reducing somewhat the effects of interactions between these factors.

Participating teams should focus on one of the following user groups (if both groups are studied, separate experiments should be run for each):

1. searchers with passive language abilities in the foreign language (i.e. that can at least roughly understand documents in that language, but cannot form accurate queries in that language without assistance). For example, a native speaker of Italian that is searching Spanish documents might be a member of this user group.
2. searchers with no useful language abilities in the foreign language. For example, a monolingual Spanish speaker that is searching German documents might be a member of this user group

How to participate

Research groups interested in joining the iCLEF 2004 task should follow these steps:

1. **Register as CLEF participants** (follow instructions in http://www.clef-campaign.org). Upon registration, every participant will receive instructions to download the appropriate document collections from the CLEF ftp site.

2. E-mail the track organizers (Doug Oard and Julio Gonzalo) indicating your wish to participate and the languages (user and document languages) that will be used in your experiment. Once registration for CLEF is confirmed, participants will receive instructions to download iCLEF question set.

3. **Formulate some hypothesis about the task**, and **design two interactive CL QA systems** intended to test your hypothesis. Usually, one of the systems is taken as a reference or baseline, and the other system is a proposed or contrastive approach. You can find examples of this methodology in previous iCLEF tracks: iCLEF 2003, iCLEF 2002 and iCLEF 2001. Ensure that both systems keep a log for post-experiment analysis that is as rich as possible. These are **examples of baseline CL QA systems** that can be used for the iCLEF 2004 task:

- A standard IR system coupled with a Machine Translation system. The user can type a question, the system translates the question automatically, retrieves relevant documents, and shows MT versions of the documents. By reading the document translations, the user may recognize possible answers, or refine the query until a suitable answer is found.
- A standard QA system coupled with a Machine Translation system. The user types a question, the system translates the question, retrieves a set of possible answers, and shows an MT version of every answer and the document that supports the answer. The user scans the answers until a seemingly correct one is found. If no answer is found, the user paraphrases the question and starts the process again.

4. Recruit subjects for the experiment; a minimum of eight subjects, more can be added in groups of eight. Make sure that the (source and target) language skills of the subjects are homogeneous. The usual setup is that the subjects are native in the question language, and have no (or very low) skills in the document language.

5. **Perform the experiment** (which takes approximately three hours per subject) and submit the results to iCLEF organizers, following the experiment design shown above.

   An experiment consists of a number of **search sessions**. A search session has three parameters: which user is searching [1-8], which question is being searched [1-16], and which system is being used [reference,contrastive]. The user has a fixed amount of time to find the answer in the document collection using the system. Once the answer is found, the user writes it down (in his own native language) in a questionnaire. Check the experiment design for details.
   Which user/question/system combinations must be carried out? We use a within-subject design like that used in early years of the TREC interactive track, but with a different number of topics and a different task. Check the experiment design for details. Overall, every user will search the full set of 16 questions (half with one system, half with the other) in an overall time (including training, questionnaries and searches) of around 3 hours.
   No formal coordination of hypotheses or comparison of systems across sites is planned for iCLEF 2004, but groups are encouraged to seek out and exploit synergies. As a first step, groups are strongly encouraged to make the focus of their planned investigations known to other track participants as soon as possible, preferably via the track listserv at iclef@listserv.uned.es. Contact julio@lsi.uned.es to join the list.

6. **Submit the results** to the track organizers following the submission format.

7. After receiving the official results, write a paper describing your experiment for the CLEF working notes and submit the paper to Carol Peters and the track organizers.

Data Provided by the Organizers

- **Document collection** : participants may choose any of the CLEF QA 2004 document languages. In each case, the documents are newswire and/or newspaper articles from major news services that were generated during 1994 and 1995. Each participant will select one of these collections that meets the needs of their chosen user group. Document collections are provided by the CLEF organization.
- **Translated documents** : The Systran machine translation system will be used at the University of Maryland to provide reference translations for some of the CLEF collections. These translations will be made available to participants through the CLEF organizers for use in their baseline system if desired. Use of these translations is not required. Translated documents will be available in the workspace for participants .
- **Questions:** 16 questions will be made for every document language. Each participating team will use the topics in the native language of the searchers. There is no restriction on the searchers' native language. Questions will be made available in the workspace for participants .
- **Questionnaires** : Questionnaires should be completed by searchers at the start of their session, after each topic, when switching systems, and at the end of their session. Participating teams are encouraged (but not required) to provide one observer per searcher if sufficient resources are available in order to maximize the value of the observational notes. Check the sample questionnaires .
- **Result format** : A standard format is provided for submitting data collected during the experiment to the organizers. This submitted data will be used as a basis for computing standard measures of effectiveness, and will be made available to any participating team upon request to facilitate more detailed cross-site comparisons. Check the submission format .

Data to be Submitted to the Organizers

Participants are encouraged to log as many details as possible about every search session. However, only a minimal information (basically, the answer provided by the user for every question/system/user combination) has to be submitted to the organizers. Check the submission format for details.

Evaluation measures

The main evaluation score for a system will be accuracy, i.e., the fraction of correct answers. Accuracy will be measured for every searcher/system pair, and then averaged over searchers to obtain a single accuracy measure for each of the two systems being compared.

The assessment will be done following the CLEF QA guidelines , except for two issues:

- When a searcher does not find an answer within the five minutes allowed per search, "NIL" will be taken as the answer for assessment purposes.

- An answer may have no supporting document. In this case, the answer will be judged correct only if it has been also found by the systems participating in the CLEF QA task, or if it can be easily retrieved from the collection by the assessors. Note that searchers must be encouraged to provide a document supporting every answer, but there is also the possibility of providing two documents or no documents in special cases (see the submission format).

The nature of any further detailed analysis is up to each site, but sites are strongly encouraged to take advantage of the experimental design and undertake exploratory data analysis to examine the patterns of correlation and interaction among factors that can be observed in the experiment, and to explore the potential for gaining additional insight through alternative evaluation measures. The computation of analysis of variance (ANOVA), where appropriate, can provide useful insights into the separate contributions of searcher, topic and system as a first step in understanding why the results of one search are different from those of another.

Schedule

Registration opens	January 15, 2004
Document release	February 2004
Question release	May 20, 2004
Submission of runs by participants	June 10, 2004
Release of individual results	July 15, 2004
Submission of papers for working notes	August 15, 2004
CLEF workshop (Bath, UK, after ECDL)	September 16-17, 2004
Document languages (choose at least one)	Dutch, French, German, Italian, Spanish, English
User languages	Unrestricted