RESEARCH AND DEVELOPMENT
ERCIM News No.40 - January 2000

CLEF - Cross-Language Evaluation Forum

by Carol Peters

A Cross-Language Evaluation Forum (CLEF), an important initiative for the evaluation of multilingual information retrieval systems, is now being launched in Europe. The activity will be sponsored by the DELOS Network of Excellence for Digital Libraries and funded by the Information Societies Technology programme of the European Commission. It will be conducted in collaboration with the US National Institute of Standards and Technology (NIST) and the TREC conference series. We present the agenda and the most important deadlines for CLEF 2000.

It has been demonstrated extensively by the Text REtrieval Conference (TREC) series that the availability of evaluation procedures can contribute significantly to the improvement of system performance. For this reason, in 1997, it was decided to include cross-language system evaluation as one of the tracks at TREC. The aim was to provide developers with an infrastructure enabling them to test and tune their systems and compare the results achieved using different cross-language strategies. From 2000 the cross-language initiative for European languages will be coordinated in Europe while TREC will focus on Asian languages. This move and the inclusion of a monolingual track for the evaluation of IR systems designed for languages other than English will help to stimulate European participation and allow us to focus on a wider range of issues.

CLEF AGENDA for 2000 - Task Description

The ultimate goal for systems for multilingual information retrieval is to offer users the opportunity to query in any language and retrieve a merged and ranked set of documents that match the query in whatever language they are stored. However, information access in multiple languages also implies an understanding of the issues involved in monolingual IR for different language types and sub-types, and many of today's applications regard cross-language retrieval between selected pairs of languages. There will thus be three evaluation tracks in CLEF 2000. Interested groups can participate in any one or in all three tracks. Newcomers to the activity may well choose to begin with the monolingual track in the first year and work up to the others in later years.

Multilingual Information Retrieval

The main task of CLEF 2000 requires searching a multilingual document collection for relevant documents in English, German, French, and Italian. Similar to the CLIR track in TREC'99, the goal is to retrieve documents from all languages, rather than just a given pair, listing the results in a merged, ranked list. Although the official languages for CLEF 2000 will be

English, French, German and Italian, it will also be possible to submit runs in which the document collection is queried in other languages. In this case, participants will be responsible for the translation of the query into their selected language. The results for such runs will be given separately.

Bilingual Information Retrieval

A cross-language task in which the query language can be either French, German or Italian but the target document collection is English will also be provided and the results will be judged. Many IR groups are now beginning to work on retrieval over pairs of languages and this will give them a chance to participate officially in the CLEF activity. Unofficial bilingual runs in which the query to the English document collection can be any language can also be submitted and will be evaluated.

Monolingual (non-English) Information Retrieval

It is often asserted that procedures for monolingual information retrieval are (almost) completely language inde-pendent. This is not however true; different languages present different problems. Methods that may be highly efficient for certain language typologies may not be so effective for others. Issues that have to be catered for include word order, morphology, diacritic characters, language variants. So far, most IR system evaluation has focussed on English. We will provide the opportunity for monolingual system testing and tuning and build up test suites in other European languages (beginning with French, German and Italian in CLEF 2000) Resources.

NIST provides a complete IR system to interested participants which currently contains simplistic German and French stemmers. This is the PRISE Test Suite.

The CLEF document collections for 2000 should consist of sets of multilingual comparable newspaper documents, from the same year, for all four languages. CLEF participants will have free access to the multilingual test suite for research purposes.

Deadlines

- Topic Release: 1 May 2000
- Receipt of results from participants: 1 July 2000
- Release of results: 15 August 2000
- Submission of paper for Working Notes: 30 August 2000
- Workshop: 21-22 Sept. 2000.

Workshop

A two-day Workshop will be held on 21-22 September in Lisbon, Portugal, immediately after ECDL 2000, the fourth European Conference on Digital Libraries. The first day will be open to all interested participants and focussed on research related issues in Multilingual Information Access. The second day will present and discuss the results of the CLEF activity and will be restricted to active CLEF participants.

Partners

NIST, Gaithersburg MD, USA (Ellen Voorhees); University of Zurich, Switzerland (Michael Hess); Social Science Information Centre, Bonn/ University of Koblenz, Germany (Jürgen Krause); CNR, Pisa, Italy (Carol Peters); Eurospider, Switzerland (Peter Schäuble).

Links:

For further information see: http://www.iei.pi.cnr.it/DELOS/CLEF/

To be included in the CLEF mailing list, please contact:

Carol Peters - IEI-CNR
Tel: +39 050 593 429
E-mail: carol@iei.pi.cnr.it

---

return to the ERCIM News 40 contents page