The Future of Evaluation for Cross-Language Information Retrieval Systems

Carol Peters[1], Martin Braschler[2], Khalid Choukri[3], Julio Gonzalo[4], Michael Kluck[5]

[1]ISTI-CNR, Area di Ricerca CNR, 56124 Pisa, Italy, carol.peters@isti.cnr.it
[2]Eurospider Information Technology AG, Zurich, Switzerland, martin.braschler@eurospider.com
[3]Evaluations and Language resources Distribution Agency, Paris, France, choukri@elda.fr
[4]Lenguajes y Sistemas Informáticos, Universidad Nacional de Educaciòn a Distancia, Madrid, Spain, julio@lsi.uned.es
[5]Informationszentrum Sozialwissenschaften (IZ), 53113 Bonn, Germany, kluck@bonn.iz-soz.de

Abstract

The objective of the Cross-Language Evaluation Forum (CLEF) is to promote research in the multilingual information access domain. In this short paper, we list the achievements of CLEF during its first four years of activity and describe how the range of tasks has been considerably expanded during this period. The aim of the paper is to demonstrate the importance of evaluation initiatives with respect to system research and development and to show how essential it is for such initiatives to keep abreast of and even anticipate the emerging needs of both system developers and application communities if they are to have a future.

1. Introduction

The Cross-Language Evaluation Forum (CLEF) has just completed its fourth year of activity. The objective of CLEF over these years has been to promote research in cross-language system development for European languages by providing an infrastructure for:

- information retrieval system testing and tuning;
- creation of an R&D community in the multilingual information access area and forum for the exchange of ideas, tools and methodologies;
- production of reusable large-scale test-suites for benchmarking purposes.

In this paper we will describe the main achievements of CLEF so far and will discuss the efforts that have been made to ensure that CLEF continues to meet the emerging needs of system developers and application communities.

2. CLEF 2000 – 2002: Multilingual Document Retrieval

Over the last decade, with the increasing globalisation of the information society, the interest in the potential of multilingual information access functionality has grown considerably. However, when the first cross-language information retrieval (CLIR) system evaluation activity began in 1997 at TREC, very little IR system testing work had been done for languages other than English and almost all existing cross-language systems were designed to handle no more than two languages: searching from query language (L1) to target language (L2). Since its beginnings, CLEF[1] has worked hard to change this situation and to promote the development of systems capable of searching over multiple languages. For this reason, each year a set of core evaluation tracks designed to test monolingual, bilingual and multilingual free text retrieval systems have been proposed. The aim has been to encourage groups to work their way up gradually from mono- to multilingual text retrieval, providing them with facilities to test and compare search and access techniques over languages and pushing them to investigate the issues involved in processing a number of languages with different characteristics. Over the years the language combinations provided have increased and the tasks offered have grown in complexity. In 2000, the main CLEF multilingual corpus consisted of approximately 360,000 newspaper and news agency documents for four languages; by 2003 it had grown to include well over 1.6 million documents and nine languages[2]. The CLEF 2003 multilingual track included a task which entailed searching a collection in eight languages.

The results in terms of participation and of the different approaches and techniques tested have been impressive. Regular CLEF participants, i.e. groups that have participated several years running, have shown improvements in performance and flexibility in advancing to more complex tasks. Much work has been done on fine-tuning for individual languages, while other efforts have concentrated on developing language-independent strategies. The issues involved in cross-language text retrieval have been investigated in depth. A discussion of these results can be found in Braschler & Peters (2004).

The evaluation environment for these tracks adopted an automatic scoring method, based on the well-known Cranfield methodology (Cleverdon, 1997). The test collection consists of a set of "topics" describing information needs and a set of documents to be searched to find those documents that satisfy the information needs. Evaluation is then done for each ranking of documents with respect to a topic by the usual computation of recall and precision. Thus, in the first years, CLEF focussed mainly on testing overall performance of off-line text retrieval systems, where good system performance is equated with good retrieval effectiveness (in terms of returning lists of documents). However, it became

[1]CLEF was launched when it was decided to move the coordination of the CLIR track at TREC to Europe. CLEF 2000 and 2001 were sponsored by the 5FP DELOS Network of Excellence for Digital Libraries; CLEF 2002 and 2003 have been funded mainly by the European Commission under the IST programme. The consortium members are ISTI-CNR, Pisa (IT); IZ-Bonn (DE); ELRA/ELDA, Paris (FR); EIT, Zurich (CH); LSI-UNED, Madrid (ES); NIST, Gaithersburg (US). CLEF 2004 will again be organised as part of the DELOS Network of Excellence, under the Sixth Framework Programme of the European Commission.

[2] The CLEF comparable corpus currently contains news documents for the same time period (1994-95) in ten languages: Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, and Swedish.

obvious that this was only one part of the CLIR problem. A multilingual system evaluation activity that meets the needs of the potential application communities must also provide facilities to investigate many other issues:

- not just document retrieval, but also targeted information location and extraction.
- not just text but also multimedia data, e.g. collections containing images or spoken documents.
- not just system performance but also wider usability issues that affect the users' ability to recognize relevant information and refine search results even if documents are written in an unfamiliar language.

In the rest of this paper we will describe how CLEF has reacted to these needs and our intentions for the future.

3. CLEF 2003-2004: Multilingual Information Access

In a SIGIR 2002 workshop[3] the question asked was whether the CLIR problem can now be considered as solved. The answers given were mixed. It was agreed that the basic technology for cross-language text retrieval systems is now in place as is clearly evidenced by numerous research papers now published[4]. But if this is so, why has this technology not been adopted by any of the large Web search engines and why do most commercial information services not offer CLIR as a standard service? The feeling was that, although there is a strong market potential, current systems are still not ready to meet the needs of the generic user. For a commercial CLIR system to be successful, it needs to be versatile, efficient when working on-line, accommodate many languages, present its results in a sufficiently user-friendly fashion, capable of handling multimedia and be conveniently coupled with fast MT engines.

At much the same time, CLEF conducted a user needs study. Two types of users were surveyed: cross-language system developers and cross-language system deployers. The main recommendations showed a lot of similarity with the opinions expressed at the workshop. Both types of users felt that CLEF should expand its range of activities further in order to meet the growing demand for testing facilities for more advanced multilingual multimedia experiments (Gonzalo et al, 2002). Not only should more languages be included in the test-suite but different types of data and tasks should be provided. In this section, we describe what has been done so far.

3.1 CLIR for Different Genres

The main CLEF collection is the multilingual comparable corpus of news documents described above. However, news media have characteristics which may not hold true for other genres: wide use of proper nouns (names and places), association of date stamps, particular style of writing and a rapid evolution of general-purpose vocabulary. Certain features may facilitate access and retrieval, others may hinder it. (Gey et al, forthcoming). For this reason, CLEF has also included a mono- and cross-language domain-specific retrieval track each year, mainly based on the GIRT corpus of structured social science data which has an associated social science thesaurus in German-English and German-Russian[5]. Since the first CLEF campaign in 2000 the GIRT corpus has been enlarged several times, and is now presented as a pseudo-parallel corpus[6], with about 150,000 identical documents in German and English. In 2002 the Amaryllis corpus[7] was also included in the domain specific task. Amaryllis consisted of French bibliographic data from all areas of science and humanities. In this way, we have provided the opportunity for developers to test their CLIR systems on domain-specific data, with domain related vocabulary and domain-specific meanings of terms[8]. Future developments may be the enlargement of the social science collection with other English, French and Russian data to make a full fledged multilingual task with domain specific data possible.

A severe problem with offering domain-specific collections in a CLEF-type evaluation campaign is that you need domain-specific experts willing to perform the relevance assessment task. Cross-lingual patent retrieval is already offered as an evaluation task by the Asian initiative, NTICIR[9], and a similar task using European language collections is on the wish-list of several CLEF groups. But the realisation of such a task depends heavily on the availability of useful data and on the willingness of experts to do the relevance assessments.

3.2 Interactive CLIR

Given a query in any source language, cross-language information retrieval has been defined as the problem of finding relevant documents written in some (different) target languages. This problem, however, represents only one of the challenges of the multilingual information access task. For instance, if a user types a Spanish query and receive a ranked list of Chinese documents, how can he/she recognize which of them are relevant for his/her purposes? How can the query be refined taking these results into account? How can information contained in documents in an unfamiliar language be exploited? These problems had not received enough attention from the research community at the time CLEF started its activities but are crucial questions for the users of a CLIR system.

The default assumptions for a cross-language search engine are that a) commercial Machine Translation (MT) systems can be used to translate documents into the native language of the user, and b) document selection and query refinement can be done using such translations. In order to challenge such (untested) assumptions, the University of Maryland, USA, and UNED, Spain, decided to organize

[3] Cross-Language Information Retrieval: A Research Roadmap Workshop at SIGIR-2002 http://ucdata.berkeley.edu/sigir-2002/

[4] See, for example, the CLEF Workshop Proceedings published by Springer: LNCS 2069, LNCS 2406, LNCS 2785.

[5] The German-English Social Science Thesaurus and the German-Russian wordlist have been provided and made available in machine readable formats by IZ-Bonn, Germany. UC Berkeley made an XML version available.

[6] The GIRT4 corpus is called pseudo-parallel because the original documents are in German and the English part consists of translations of these German documents into English; the English part is actually considerably smaller than the German.

[7] The Amaryllis collection was made available by the Institut de l'Information Scientifique et Technique (INIST), France.

[8] Cf. Kluck & Gey, 2001.

[9] NTCIR (NII-NACSIS Test Collection for IR Systems): http://research.nii.ac.jp/ntcir/

an interactive track – *i*CLEF – within CLEF. The goal was comparative studies of user interaction issues in CLIR.

The first pilot track of *i*CLEF was held in 2001 (Oard and Gonzalo, 2002), and focused on document selection questions: Is MT the only option available? Can it be substituted by simpler, faster methods without degradation of relevance judgements? Are there preferable alternatives? The *i*CLEF experimental design to test these issues combined insights from the interactive TREC track (Over, 2001) with the distributed multilingual assessment process of CLEF. The experience led to non-trivial empirical conclusions; for instance, full MT performed better than word-by-word translation, but cross-language pseudo-summaries (allowing faster relevance judgements without loss of precision) outperformed full MT. The success of this experiment led to the *i*CLEF track being included as a regular event in CLEF.

*i*CLEF tracks in CLEF 2002 and 2003 studied support mechanisms for interactive query formulation and refinement (Gonzalo & Oard, 2003; Oard & Gonzalo, 2004), and included experiments such as user-assisted term translation (via inverse dictionaries, translation definitions, etc), which was shown to outperform automatic query translation, and user-assisted query reformulation by selection of relevant noun-phrases, which was shown to outperform user-assisted term translation.

Overall, *i*CLEF experiences have involved around five hundred interactive cross-language searches in several languages (English, Spanish, Finnish, German and Swedish), constituting the largest set of empirical data about multilingual information access known to us.

3.3 Multilingual Question Answering

Question Answering (QA) systems accept natural language questions and retrieve answers, rather than documents, from a text collection. Since the first QA TREC track (Voorhees, 1999), this task has received a growing attention from the IR and natural language processing research communities, and is currently the largest track (in number of participants) at TREC. Multilinguality has been identified as one of the major challenges for QA systems (Maybury, 2002). However, little is yet known about cross-language QA, i.e., when the question is not written in the same language as the documents. It is certainly a harder problem (from the point of view of translinguality) than CLIR: while, for document retrieval purposes, finding appropriate candidate translations for the query words can lead to good performance, questions demand a careful linguistic analysis (in the source language) which is not trivially translated into the target language; and a (possibly noisy) machine-translated question may produce errors in the linguistic analysis.

In order to study these issues, in 2003, three research groups from Italy, Spain and the Netherlands (ITC-irst, UNED and ILLC) organized a pilot Cross-Language QA track under the auspices of CLEF. This track evaluated monolingual QA systems in Dutch, Italian and Spanish, and cross-language QA systems searching an English target collection with Spanish, Italian, Dutch, French or German questions (Magnini et al., 2003). Although only limited comparisons between systems could be made (there was only one participating system per monolingual task) and just 6 groups participated in the bilingual experiment, the experience was considered a success, generating the first cross-language QA test suite and attracting a lot of interest from CLEF participants.

3.4 Cross-Language Retrieval in Image Collections

A cross-language image retrieval task, known as ImageCLEF, was introduced into CLEF in 2003 (Clough & Sanderson, 2004). This was probably the first time that the research community began to think seriously about the issues involved in retrieval from an image collection when the user queries are expressed in a language different from that of the collection. A mixture of language dependent and language independent, cross-language and cross-cultural factors are involved, according to whether retrieval is based on low-level features derived from an image, or on the associated caption, or on a combination of both. ImageCLEF aims to provide the necessary collection(s) and framework in which to analyse the link between image and text and promote the discovery of alternate methods for cross-language image retrieval. However, research in cross-language image retrieval is not just of academic interest, there is a strong commercial potential as organisations with large image collections would be able to offer the same collections to a wider range of users with differing language backgrounds.

The goal is thus to provide a test bed that can be used to evaluate different retrieval methods and to analyse user behaviour during the search process, eg query formulation in both cross-language and visual environments, iterative searching and query reformulation.

The first ImageCLEF was set up as a pilot experiment; just four groups participated in an ad hoc retrieval task. Participants were free to use either content-based or text-based retrieval methods, relevance feedback and any translation method. The search requests were provided in Dutch, French, German, Italian and Spanish and consisted of both visual examples and text descriptions of the user need. The image collection of approximately 30,000 historical photographs of Scotland complete with short captions was made available by St Andrews University Library. Although all groups chose to use the information derived from captions only, there was much interest in the potential of this exercise at the CLEF 2003 workshop and it was decided to expand it in CLEF 2004, adding another collection (medical images complete with case notes in French and English) and introducing a task in which participants are obliged to use both visual and text data in their search.

3.5 Cross-Language Spoken Document Retrieval

CLEF began to pay attention to the issues involved in cross-language spoken document retrieval (CL-SDR) in 2002 when a pilot experiment was organised within the DELOS Network of Excellence by two groups (ITC-irst and University of Exeter) and the results were reported at the CLEF workshop that year (Jones & Federico, 2003). The experiment was continued on a slightly larger scale as a regular track in CLEF 2003 with four participating groups. As resources are limited so far the track has used data used in the TREC 8 and 9 English monolingual SDR

tracks, kindly made available by NIST and the results are closer to a benchmark rather than a real evaluation. The TREC collections have been extended to a CL-SDR task by manually translating the short topics into five European languages: Dutch, French, German, Italian and Spanish. In 2003 the track aimed at evaluating CLIR systems on noisy automatic transcripts of spoken documents with known story boundaries. The results of the experiments showed that, as expected, bilingual performance was lower for all participants than the comparative English monolingual run. However, the degree of degraded performance was shown to depend on the translation resources used. It was also shown that it can be effective to use different indexing units for monolingual and bilingual retrieval on the data set (Federico & Jones, 2004).

4. Future Prospects

It is generally agreed that evaluation campaigns can play an active role in advancing system development. The goal of CLEF at the moment is to attempt to narrow the gap between the R&D community and application world. We have already begun to do this by including activities that investigate different kinds of user-system interaction and that test system performance on collections that are not just text-oriented but consider the needs of other media.

CLEF 2004 is continuing in this direction. Less attention is being given this year to the ad hoc text retrieval tasks and more attention to the newer activities. In particular, ImageCLEF has been expanded with the addition of new tasks and a new collection, and a major focus will be on user satisfaction issues. Both the multilingual question answering track and ImageCLEF will be collaborating with iCLEF in order to introduce interactive tasks in their tracks. At the same time, we are beginning to discuss our plans for the future. The question is: what does the user most need? Due to the lack of space we just mention one idea here: multilingual Web search testing.

A Web track has already been offered by TREC and NTCIR but in a monolingual context. As was noted at the SIGIR workshop, current web search engines do not offer truly multilingual search functionality. We believe CLEF could provide a real service to the community by providing a multilingual Web test collection and testing systems with respect to different issues, eg results presentation, user interaction. Tasks that could be offered include finding a known page (from a description of it or parts of the name), identifying the geographical location of a given company, classifying results by topic. Unfortunately, the building of a Web corpus for multilingual access is not a trivial task; the basic snapshot data of the Web and the evaluation metrics would have to be defined very carefully (see Gurrin & Smeaton, 2003 and Kando, 2004). A multilingual Web corpus would have to include carefully calculated representative samples for less commonly used languages plus random samples for the most dominant ones. Questions of intellectual property rights would also have to be considered. CLEF is considering following the TREC example, and using urls and pages taken from the government domain in many European countries to build a multilingual, comparable web corpus.

To find out more about CLEF past and future activities, see: http://www.clef-campaign.org.

5. References

Braschler, M. & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, 7(1-2) pp 7-31.

Cleverdon, C. (1977). The Cranfield Tests on Index Language Devices. In: K. Spärck-Jones and P. Willett, eds. Readings in Information Retrieval, Morgan Kaufmann, 1997. pp 47-59.

Clough, P.D. & Sanderson, M. (2004). The CLEF 2003 Cross Language Image Retrieval Track. In CLEF 2003 Proceedings, LNCS, Springer, forthcoming.

Federico, M. & Jones, G.J.F. (2004). The CLEF 2003 Cross-Language Spoken Document Retrieval Track. In CLEF 2003 Proceedings, LNCS, Springer, forthcoming.

Gey, F., Kando, N., Peters, C. (2004). Cross-Language Information Retrieval: the Way Ahead. In Information Processing and Management. Special issue on CLIR, forthcoming.

Gonzalo, J. & Oard, D. (2003). The CLEF 2002 Interactive Track. In CLEF 2002 Proceedings, Springer-Verlag LNCS 2785, pp 372-382.

Gonzalo, J., Verdejo, F., Peñas, A. & Peters, C. (2002). User Needs: Del. 1.1.1, CLEF project: http://clef.iei.pi.cnr.it:2002/deliv_avail_to_public/Del111.pdf

Gurrin, C. & Smeaton, A. F. (2003). Improving Evaluation of Web Search Systems. In Sebastiani, F. (ed.) Advances in Information Retrieval. Springer LNCS 2633, pp 25-40

Jones, G.J.F. & Federico, M. (2003). CLEF 2002 Cross-Language Spoken Document Retrieval Pilot Track Report. In Proceedings of CLEF 2002, Springer-Verlag LNCS 2785, pp 446-457.

Kando, N. (2004) Evaluation of Information Access Technologies at the NTCIR Workshop. In CLEF 2003 Proceedings, LNCS, Springer, forthcoming.

Kluck, M. & Gey, F. C. (2001): The Domain-Specific Task of CLEF – Specific Evaluation Strategies in Cross-Language Information Retrieval. In Proceedings of CLEF 2001, Springer-Verlag LNCS 2406, pp 48-56

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. & de Rijke, M. (2004). The Multiple Language Question Answering Track at CLEF 2003. In CLEF 2003 Proceedings, LNCS, Springer, forthcoming.

Maybury, M. (2002) Toward a Question Answering Roadmap, MITRE Tech. Report.

Oard, D. & Gonzalo, J. (2002). The CLEF 2001 Interactive Track. In Proceedings of CLEF 2001, Springer-Verlag LNCS 2406. pp 308-319.

Oard, D. & Gonzalo, J. (2004). The CLEF 2003 Interactive Track. In CLEF 2003 Proceedings, LNCS, Springer, forthcoming.

Over, P. (2001). The TREC Interactive Track: an annotated bibliography. Information Processing and Management, 37(3) pp 369-381.

Voorhees, E. (1999). The TREC-8 Question Answering Track Report. In Proceedings of TREC8, NIST special publication 500-246, pp. 77-82.