The Impact of Evaluation on Multilingual Information Retrieval System Development

Carol Peters

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche
Via Moruzzi, 1, 56124 Pisa, Italy
carol.peters@isti.cnr.it

Abstract

The Cross-Language Evaluation Forum (CLEF) promotes research into the development of truly multilingual systems capable of retrieving relevant information from collections in many languages and in mixed media. The paper discusses some of the main results achieved in the first six years of activity.

1. Introduction

The objective of the Cross-Language Evaluation Forum (CLEF) is to promote research in the field of multilingual system development. CLEF thus organizes annual evaluation campaigns in which a series of tracks designed to test different aspects of mono- and cross-language information retrieval (IR) are offered. The intention is to encourage experimentation with all kinds of multilingual information access – from the development of systems for monolingual retrieval operating on many languages to the implementation of complete multilingual multimedia search services. This has been achieved by offering an increasingly complex and varied set of evaluation tasks over the years. The aim is not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems. In the following sections, we briefly describe the organization of the CLEF campaigns and (some of) the results achieved. Proposals for future directions are given in the conclusions.

2. CLEF Campaigns 2000-2005

CLEF actually began life in 1997 as a track for cross-language information retrieval (CLIR) within TREC, the well-known Text REtrieval Conference series sponsored in the US by NIST and DARPA[1]. At that time, almost all existing cross-language systems were designed for text retrieval and handled only two languages, searching from query language to target language. In addition, for most of these systems one of the two languages was English. Thus, three years later, when the coordination of this activity was moved to Europe and CLEF was launched as an independent initiative[2], our primary goals were the promotion of system testing and evaluation for European languages other than English and the development of truly multilingual retrieval systems, capable of retrieving relevant information from collections in many languages and in mixed media.

[1] See http://trec.nist.gov/

[2] The first CLEF evaluation campaign was held in 2000 and culminated in a workshop in Lisbon, Portugal, in September of that year. CLEF is currently an activity of the DELOS Network of Excellence under the Sixth Framework programme of the European Commission. For more information, see http://www.clef-campaign.org/.

CLEF 2000
• mono-, bi- and multilingual textual document retrieval (Ad Hoc)
• mono- and cross-language information on structured scientific data (Domain-Specific)
CLEF 2001
• interactive cross-language retrieval (iCLEF)
CLEF 2002
• cross-language spoken document retrieval (CL-SR)
CLEF 2003
• multiple language question answering (QA@CLEF)
• cross-language retrieval in image collections (ImageCLEF
CLEF 2005
• multilingual retrieval of Web documents (WebCLEF)
• cross-language geographical retrieval (GeoCLEF)

Table 1: CLEF 2000 – 2005: increase in tracks

The CLEF evaluation campaigns have been designed in order to work towards these goals. Tracks are proposed to examine particular areas of cross-language IR and are subdivided into tasks, which can vary from year to year, according to the specific aspects of system performance to be tested. Table 1 shows how the number of tracks has been extended since 2000, to reach a total of eight in 2005. Figure 1 shows how the focus of CLEF has shifted from textual document retrieval to encompass cross-language retrieval for mixed media (speech and image) and targeted information extraction in a multilingual context (question answering and geographic retrieval).

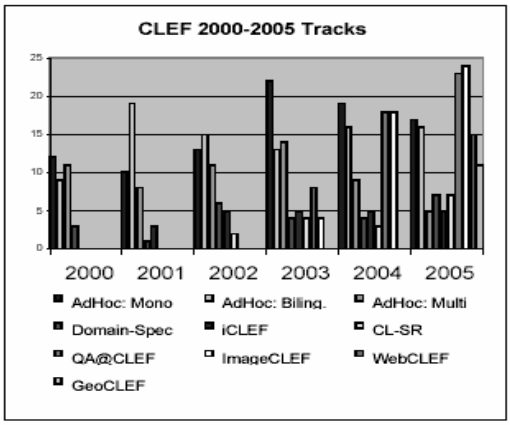

Figure 1: CLEF 2000-2005 Shift in focus

3. Test Collections

CLEF campaigns adopt a comparative evaluation approach in which system performance is measured using appropriate test suites (Cleverdon, 1997). These consist of sets of sample query statements often called "topics", document collections, and relevance judgments determining the set of relevant documents in a collection for a given query statement. Seven different document collections were used to build the test sets in CLEF 2005:

- a multilingual comparable corpus of more than 2 million news docs in twelve European languages
- social science databases in English, German and Russian
- a historical photographic archive
- a radiological medical database with case notes in French and English
- an English/German for database automatic medical image annotation
- a collection of spontaneous conversational speech derived from the Shoah archives
- a multilingual collection of about 2M web pages crawled from European governmental sites.

For each collection, appropriate sets of search requests and associated relevance assessments have been built. These test suites form extremely valuable and reusable resources. They are created according to rigorous guidelines and are tested to confirm their stability. It is our intention to make them publicly available via the ELDA catalogue. ELDA representatives are currently finalizing agreements with the data providers for this purpose.

4. Results

In this section, we outline some of the principal results achieved by CLEF with respect to the main goal of promoting the development of multilingual/multimedia information retrieval systems. For complete documentation on individual CLEF experiments and results, track by track and year by year, see the on-line CLEF Working Notes at http://www.clef-campaign.org/.

4.1 Cross-language Text Retrieval

CLEF has tried to encourage groups to work their way up gradually from mono- to true multilingual text retrieval by providing them with facilities to test and compare search and access techniques over many languages, pushing them to investigate the issues involved in processing a growing number of languages with different characteristics.

	Monolingual	Bilingual	Multilingual
CLEF2000	DE;FR;IT	X→EN	X→DE;EN;FR;IT
CLEF2001	DE;ES;FR;IT NL	X→EN, X→NL	X→DE;EN;ES; FR;IT
CLEF2002	DE;ES;FI;FR IT;NL;SV	X→DE;ES;FI;FR IT;NL;SV X→EN(newcomer)	X→DE;EN;ES; FR;IT
CLEF2003	DE;ES;FI;FR IT;NL;RU;SV	IT→ES;DE→IT FR→NL;FI→DE X→RU;X→EN	X→DE;EN;ES;FR X→DE;EN;ES;FI FR;IT;NL;SV
CLEF2004	FI;FR;RU;PT	ES/FR/IT/RU→FI DE/FI/NL/SV→FR X→RU;X→EN	X→FI;FR;RU;PT
CLEF2005	BG;FR;HU;PT	X→ BG;FR;HU;PT	Multi8 2yrson Multi8 merge

BG=Bulgarian;DE=German;EN=English;ES=Spanish;FI=Finnish;FR=French
HU=Hungarian;IT=Italian;NL=Dutch;PT=Portuguese;RU=Russian;SV=Swedish

Table 2: CLEF 2000 – 2005 Ad-Hoc Tasks

As can be seen from Table 2, we have now created ad-hoc cross-language test collections for twelve European languages. Over the years the language combinations have increased and the tasks offered have grown in complexity until, in CLEF 2003, the multilingual track included a task which entailed searching a collection in 8 languages, selected to cover a range of language typologies and linguistic features (Multi-8). We also encouraged system testing with uncommon language pairs (e.g. German to Italian or French to Dutch) in both 2003 and 2004. Instead, the multilingual task in CLEF 2005 was designed to focus on a particular aspect of the multilingual retrieval problem: the merging of results over different languages and collections.

4.1.1 Performance Improvement

Groups submitting results over several years have shown 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. However, an important question is whether we can demonstrate improvements in system performance. As test collections and tasks vary over years, such improvements are not easy to document. For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. Some findings are reported here::

In 1997, at TREC-6, the best cross-language text retrieval systems had the following results:

- EN→FR: 49% of best monolingual French system
- EN→DE: 64% of best monolingual German system

In 2002, at CLEF, where there was no restriction on topic and target language, the best systems gave:

- EN→FR: 83,4% of best monolingual French system
- EN→DE: 85,6% of best monolingual German system

CLEF 2003 enforced the use of "unusual" language pairs, with the following impressive results:

- IT→ES: 83% of best monolingual Spanish IR system
- DE→IT: 87% of best monolingual Italian IR system
- FR→NL: 82% of best monolingual Dutch IR system

In CLEF 2005, where we introduced two new languages, we found:

- X→FR: 85% of best monolingual French system
- X→PT: 88% of best monolingual Portuguese system
- X→BG: 74% of best monolingual Bulgarian system
- X→HU: 73% of best monolingual Hungarian system

From these figures, we can see that there is a general trend of improvement in bilingual system performance which tends to stabilize. With languages for which testing has gone on for several years, there is usually little variation in performance between the best groups, whereas for "new" languages where there has been little CLIR system testing, there is normally room for improvement (see the examples of Bulgarian and Hungarian in CLEF 2005).

In CLEF 2005 we attempted to reuse the Multi-8 test collection created in CLEF 2003 to see whether a similar improvement in multilingual system performance could be measured, and also to examine the results merging problem. Unfortunately, there was not a numerous participation in this task and the results obtained are only indicative. However, we can report that the top performing submissions to both the multilingual 2-Years-On and the merging tasks were better than the best submission to the CLEF 2003 Multi-8 task.

Summing up, we find that, over the years, CLEF participants learn from each other and build up a collective knowhow. Thus, as time passes, we see a convergence of techniques and results with very little statistical difference between the best systems. We have observed that the best systems are a result of careful tuning of every component, and of combining different algorithms and information sources for every subtask (see Braschler & Peters, 2004).

4.2 Cross-language Information Extraction

For many years, IR has concentrated on document retrieval. However, users often want specific answers rather than all the information that is to be found on a given topic. For this reason, information extraction systems have been given much attention. In 2003, CLEF introduced a cross-language question-answering track thus stimulating the development of some of the very first multilingual QA systems. CLEF 2005 ran pilot experiments in cross-language geographic IR. It is too early to have significant results for this activity as yet.

4.2.1 Multilingual Question Answering

Question answering systems have been evaluated for many years at TREC and the track has evolved over the years to offer increasingly difficult tasks. However, multilinguality has never been taken into consideration. As QA techniques are mainly based on natural language processing tools and resources, we felt that it was important to fill this gap in CLEF. The aim of the track is to encourage testing on languages other than English, to check and/or improve the portability of technologies implemented in English QA systems, and to force the QA community to design real multilingual systems. The QA@CLEF campaign in 2005 was the result of experience acquired during the two previous years and proved very popular. 24 participating groups submitted mono- and cross-language runs for nine target collections. In these three years, performance for both mono- and cross-language systems has shown improvement, with the best non-English systems in 2005 obtaining very similar results to those of TREC, and the best bilingual systems obtaining a performance of approximately 60% of monolingual results. From a comparison of approaches, we see that most systems pre-process the document collection, adopting linguistic processors and language resources such as POS-taggers, named entity recognizers, WordNet, gazetteers. Many systems adopt a deep parsing strategy while only a few use any logical representation.

4.3 Cross-language Multimedia Retrieval

The current growth of multilingual digital material in a combination of different media (e.g. image, speech, video) means that there is an increasing interest in systems capable of automatically accessing the information available in these archives. For this reason, CLEF supported a preliminary investigation aimed at evaluating systems for cross-language spoken document retrieval in 2002 and in 2003 introduced a track for cross-language retrieval on image collections.

4.3.1 Cross-Language Speech Retrieval

The 2005 Cross-Language Speech Retrieval (CL-SR) track followed two years of experimentation with cross-language retrieval of broadcast news in CLEF 2003 and CLEF 2004. In 2005 the track focused on spontaneous speech retrieval over languages. Spontaneous speech is considerably more challenging for the Automatic Speech Recognition (ASR) techniques on which fully-automatic content-based search systems are based. Recent advances in ASR have made it possible to contemplate the design of systems that would provide a useful degree of support for searching large collections of spontaneous conversational speech, but no representative test collection that could be used to support the development of such systems has been widely available for research use. The principal goal of the CLEF-2005 CL-SR track was thus to create such a test collection. Additional goals included benchmarking the present state of the art for ranked retrieval of spontaneous conversational speech and fostering interaction among a community of researchers with interest in that challenge. The collection used was a set of interviews with Holocaust survivors, extracted from the Shoah archives.

Just seven teams from four countries participated in this track in 2005. A reusable test collection for searching spontaneous conversational English speech using queries in five languages (Czech, English, French, German and Spanish) was built and includes speech recognition for spoken words, manually and automatically assigned controlled vocabulary descriptors for concepts, dates and locations, manually assigned person names, and hand-written segment summaries. The 2006 CL-SR track will extend this collection to include additional English speech (about 900 hours), additional resources (word lattices and more accurate speech recognition), and a no-boundary evaluation condition. A second test collection containing at least 500 hours of Czech speech will also be created.

4.3.2 ImageCLEF

The ImageCLEF retrieval benchmark aims at evaluating image retrieval from multilingual document collections. Images by their very nature are language independent, but are often accompanied by semantically related texts (e.g. captions or metadata). Images can then be retrieved using primitive features based on pixels which form the contents of an image (e.g. using a visual exemplar), abstracted features expressed through text, or a combination of both. The language used to express the associated texts or textual queries should not affect retrieval, i.e. an image with a caption written in English should be searchable in languages other than English.

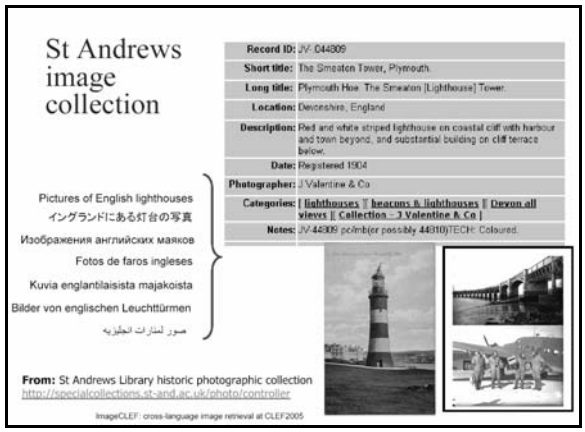

Figure 2: ImageCLEF 2005 historical photo request

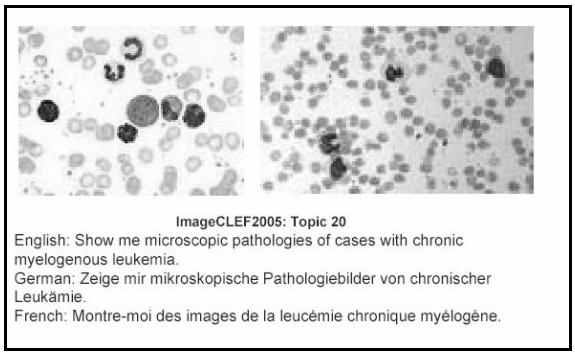

Figure 3: ImageCLEF 2005 medical task request

A major goal of ImageCLEF is to investigate the effectiveness of combining text and image for retrieval and to promote the exchange of ideas which may help improve the performance of future image retrieval systems. Participants are provided with image collections, representative search requests (expressed by both image and text) and relevance judgements indicating which images are relevant to each search request. ImageCLEF 2005 provided tasks for system-centered evaluation of retrieval systems in two domains: historic photographs and medical images. These domains offer realistic (and different) scenarios in which to test the performance of image retrieval systems and present different challenges and problems to participants. Figures 2 and 3 show search requests for these scenarios. In both tasks, the best results were obtained by systems that combined text and content-based retrieval mechanisms.

ImageCLEF is important because more research into multimodal retrieval, combining text and visual features and catering also for multilinguality is needed. For this reason, it is not surprising that this track was very popular in CLEF2005 with a large participation, and is expected to be even more so in 2006.

5. Conclusions and Future Directions

The results achieved by CLEF in the first six years of activity are impressive. We can summarise them in the following main points:

- documented improvement in system performance for cross-language text retrieval systems
- quantitative and qualitative evidence with respect to best practice in cross-language system development
- R&D activity in new areas such as cross-language question answering, multilingual retrieval for mixed media, and cross-language geographic information retrieval
- creation of important, reusable test collections for system benchmarking
- building of a strong, multidisciplinary research community.

Furthermore, CLEF evaluations have provided qualitative and quantitative evidence along the years as to which methods give the best results in certain key areas, such as multilingual indexing, query translation, resolution of translation ambiguity, results merging.

However, although CLEF has done much to promote the development of multilingual IR systems, so far the focus has been on building and testing research prototypes rather than developing fully operational systems. There is still a considerable gap between the research and the application communities and, despite the strong demand for and interest in multilingual IR functionality, there are still very few commercially viable systems on offer. The challenge that CLEF must face in the near future is how to best transfer the research results to the market place. CLEF 2006 is taking a first step in this direction with the organisation of a real time exercise as part of the question-answering track. The aim is to measure system performance not only according to the accuracy of the replies but also with respect to the response times.

6. Acknowledgments

We gratefully acknowledge the support of the data providers. Without their contribution, this evaluation activity would be impossible:

- The LA Times, for the American English data collection
- SMG Newspapers for the British English data
- Le Monde S.A. and ELDA: Evaluations and Language resources Distribution Agency, for the French data
- Frankfurter Rundschau, Druck und Verlagshaus Frankfurt am Main; Der Spiegel, Spiegel Verlag, Hamburg, for the German newspaper collections
- InformationsZentrum Sozialwissen-schaften, Bonn, for the GIRT database
- SocioNet for the Russian Social Science Corpora
- Hypersystems, Torino and La Stampa, for Italian data
- Agencia EFE S.A. for the Spanish data
- NRC Handelsblad, Algemeen Dagblad and PCM Landelijke dagbladen/Het Parool for Dutch data
- Aamulehti Oyj & Sanoma Osakeyhtiö for Finnish data
- Russika-Izvestia for the Russian newspaper data
- Público, Portugal, and Linguateca for Portuguese data
- Folha, Brazil, and Linguateca for Brazilian newspapers
- Tidningarnas Telegrambyrå for the Swedish data
- Schweizerische Depeschenagentur, Switzerland, for French, German and Italian Swiss news agency data
- Ringier Kiadoi Rt and Research Inst. for Linguistics, Hungarian Acad. Sci. for Hungarian newspapers
- Sega AD, Sofia; Standart Nyuz AD, Sofia, and the BulTreeBank Project, Bulgarian Acad. Sci, for the Bulgarian newspaper documents
- St Andrews U. Library for historic photographic data
- U. and Uni Hospitals, Geneva, Switzerland and Oregon Health and Science U. for the ImageCLEFmed Radiological Medical Database
- Aachen University of Technology (RWTH), Germany for the IRMA database of annotated medical images
- The Survivors of the Shoah Visual History Foundation, and IBM for the Malach spoken document collection

1. References

Braschler, M., Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, Vol.7 (1-2), 5-29.

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