From Research to Application in Multilingual Information Access: the Contribution of Evaluation

Carol Peters[1], Martin Braschler[2], Giorgio Di Nunzio[3], Nicola Ferro[3], Julio Gonzalo[4], Mark Sanderson[5]

[1]Istituto di Scienza e Tecnologie dell'Informazione, Via Moruzzi, 1, 56124 Pisa, Italy
[2]Zurich University of Applied Sciences, Switzerland
[3]University of Padua, Italy
[4]Universidad Nacional de Educación a Distancia, Madrid, Spain
[5]University of Sheffield, United Kingdom
carol.peters@isti.cnr.it, martin.braschler@zhwin.ch, dinunzio|ferro@dei.unipd.it, julio@lsi.uned.es,
m.sanderson@sheffield.ac.uk

Abstract

The importance of evaluation in promoting research and development in the information retrieval and natural language processing domains has long been recognised but is this sufficient? In many areas there is still a considerable gap between the results achieved by the research community and their implementation in commercial applications. This is particularly true for the cross-language or multilingual retrieval areas. Despite the strong demand for and interest in multilingual IR functionality, there are still very few operational systems on offer. The Cross Language Evaluation Forum (CLEF) is now taking steps aimed at changing this situation. The paper provides a critical assessment of the main results achieved by CLEF so far and discusses plans now underway to extend its activities in order to have a more direct impact on the application sector.

1. Introduction

A major result of the global information society is that the Internet has become the primary source of information of all types for much of the population. Every year the proportion of English content is decreasing as information is increasingly being made available in more of the world's languages[1]. There is thus a strongly felt need for efficient multilingual information access (MLIA) systems that allow users to search document collections in multiple languages and retrieve relevant information in a form that is understandable to them, even when they have little or no linguistic competence in the target languages concerned. MLIA systems must handle all document types (e.g. text, image, speech and video) and through tasks like summarization, filtering and question-answering, provide high-level and innovative access to document collections, thus enabling genuine information retrieval as opposed to traditional document retrieval. They must be capable of finding the information that the user requires but also of presenting it in a way that is easily understandable and reusable. The Cross Language Evaluation Forum (CLEF) was set up in 2000[2] to stimulate the development of such systems.

2. Cross Language Evaluation Forum

The main objectives of CLEF are to stimulate the development of mono- and multilingual information retrieval systems for European languages. These objectives are realised through the organisation of annual evaluation campaigns and workshops. The scope of CLEF has gradually expanded over the years. While in the early years, the main interest was in textual document retrieval, the focus is now diversified to include different kinds of text retrieval across languages (ie not just document retrieval but question answering and geographic IR as well) and on different kinds of media (ie not just plain text but collections also containing images and speech). The aim has been to encourage experimentation with all kinds of multilingual information access – from the development of systems for monolingual information retrieval (IR) operating on many languages to the implementation of complete multilingual multimedia search services. The goal has been 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. CLEF has given research groups access to standardized testbeds, allowing evaluation of their approaches and comparison across various types of multilingual systems[3].

2.1 CLEF Evaluation Infrastructure

The Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) is a sophisticated evaluation infrastructure which provides a set of tools capable of managing high-level tasks, such as topic creation, experiment submission, pool assessment, relevance assessment, statistical analysis on the experiments, etc. (Di Nunzio & Ferro, 2006). DIRECT has been successfully employed in the last three CLEF campaigns (2005, 2006, and 2007) evaluation campaigns with the following goals:

- to approach evaluation initiatives in a new way, able to adopt a data curation approach to experimental evaluation;
- to better model the experimental data and the metadata associated to them;

[1] In 2006 English represented ca 30% of Internet content; the next most represented language is Chinese at nearly 14%, see http://www.internetworldstats.com/stats.htm

[2] See www.clef-campaign.org

[3] Over the years, CLEF has created test collections for evaluation of text, image and speech retrieval systems working in the multilingual context, with target document collections in thirteen languages.

- to provide participants with a uniform way of performing statistical analysis on their experiments;
- to support the long term preservation, lineage, curation and enrichment of the experimental data.

2.2 Results

This activity has led to the creation of important, reusable test collections for system benchmarking [4] and has provided valuable quantitative and qualitative evidence with respect to best practice in cross-language system development. For example, CLEF evaluations have provided 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 (Braschler & Peters, 2004).

There is also substantial proof of significant increase in retrieval effectiveness in multilingual settings by the systems of CLEF participants. The first evaluation of cross-language systems on European languages actually began in 1996 at the Text REtrieval Conference (TREC) series organised in the United States, testing on English-French and English-German retrieval. This activity continued for three years at TREC, including also Italian as an additional language, and was then moved to Europe in 2000 with the launch of CLEF. (Braschler, 2004) provides a comparison between effectiveness scores from the 1996 TREC-6 campaign and the CLEF 2003 campaign in which retrieval tasks were offered for eight European languages. While in 1996 systems were performing at about 50%-60% of monolingual effectiveness for multilingual settings, that figure had risen to 80%-85% by 2003 for languages that had been part of multiple evaluation campaigns.

The main results of the CLEF activity over the years can be summarised in the following points:

- Stimulation of research activity in new, previously unexplored areas, such as cross-language question answering, image and geographic information retrieval
- Study and implementation of evaluation methodologies for diverse types of cross-language IR systems
- Documented improvement in system performance for cross-language text retrieval systems
- Creation of a large set of empirical data about multilingual information access from the user perspective
- Quantitative and qualitative evidence with respect to best practice in cross-language system development
- Creation of important, reusable test collections for system benchmarking
- Building of a strong, multidisciplinary research community.

However, although CLEF has done much to promote the development of multilingual IR systems, the focus has been on building and testing research prototypes rather than developing fully operational systems. The challenge that must now be faced is how to best transfer these research results to the market place.

3. From R&D to Technology Transfer

So far, CLEF has been a forum where researchers can perform experiments, discuss results and exchange ideas; most of the results have been published but the extensive CLEF-related literature is mainly intended for the academic community. Contacts with interested application communities have been notably lacking.

In fact, it should be observed that evaluation campaigns also have their limitations. They tend to focus on aspects of system performance that can be measured easily in an objective setting (e.g. precision and recall) and to ignore others that are equally important for overall system development. Thus, while CLEF, much attention has been paid to improving performance in terms of the ranking of results through the refining of query expansion procedures, term weighting schemes, algorithms for the merging of results, equally important criteria of speed, stability, usability have been mainly ignored. Also for any MLIA system, the results must be presented in an understandable and useful fashion. The user interface implementation thus needs to be studied very carefully according to the particular user profile. Such aspects tend to be neglected in traditional evaluation campaigns.

We have thus decided to launch a new activity which aims at building on and extending the results already achieved by CLEF. This activity, called TrebleCLEF [5], will stimulate the development of operational MLIA systems rather than research prototypes.

4. TrebleCLEF

TrebleCLEF intends to promote research, development, implementation and industrial take-up of multilingual, multimodal information access functionality in the following ways:

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to stimulate R&D to meet the requirements of the user and application communities, with particular focus on the following key areas:
  - user modeling, e.g. what are the requirements of different classes of users when querying multilingual information sources;
  - language-specific experimentation, e.g. looking at differences across languages in order to derive best practices for each language, best practices for the development of system components and best practices for MLIA systems as a whole;
  - results presentation, e.g. how can results be presented in the most useful and comprehensible way to the user.
- by constituting a scientific forum for the MLIA community of researchers enabling them to meet and discuss results, emerging trends, new directions:
  - providing a scientific digital library to manage accessible the scientific data and experiments produced during the course of an evaluation

[4] The 2000-2003 test collections are now publicly available on the ELDA catalog, see www.elda.org.

[5] TrebleCLEF is a 7FP Coordination Action under the IST programme; it began activity in January 2008. The Consortium is composed of five academic partners and two important centres: ISTI-CNR, Italy; University of Padua, Italy, University of Sheffield, UK; Zurich University of Applied Sciences, Switzerland; UNED, Spain; CELCT, Italy, ELDA, France. See www.trebleclef.eu.

campaign. This library would also provide tools for analyzing, comparing, and citing the scientific data of an evaluation campaign, as well as curating, preserving, annotating, enriching, and promoting the re-use of them;

- by acting as a virtual centre of competence providing a central reference point for anyone interested in studying or implementing MLIA functionality and encouraging the dissemination of information:
  - making publicly available sets of guidelines on best practices in MLIA (e.g. what stemmer to use, what stop list, what translation resources, how best to evaluate, etc., depending on the application requirements);
  - making tools and resources used in the evaluation campaigns freely available to a wider public whenever possible; otherwise providing links to where they can be acquired;
  - organising workshops, and/or tutorials and training sessions.

4.1 A Digital Library for Evaluation

The experimental data produced during an evaluation campaign are valuable scientific data, and as a consequence, should be archived, enriched, and curated in order to ensure their future accessibility and re-use (Agosti, Di Nunzio & Ferro, 2007). In TrebleCLEF the DIRECT system already used in the CLEF evaluation campaigns will be extended in order to provide a coherent and uniform way for preserving and accessing the scientific data resulting from the evaluation activity, and maximizing the benefits of information technology for better access to and easier use of scientific knowledge.

As pointed out by (Agosti, Di Nunzio, & Ferro, 2007), a DLS is “the natural choice for managing, making accessible, citing, curating, enriching, and preserving all the information resources produced during an evaluation campaign” since it provides a more mature way of dealing with the scientific data produced during the IR experimental evaluation. Much care must be given to the design and development of the user interface of DLS of this type, since it needs to be able to support high-level cognitive tasks and the investigation and understanding of the experimental results (Dussin & Ferro, 2007).

Scientific data, their enrichment and interpretation are essential components of scientific research. The so-called “Cranfield methodology”(Cleverdon, 1967), which is the paradigm usually followed by an evaluation campaign, traces how these scientific data have to be produced, while the statistical analysis of experiments provide the means for further elaborating and interpreting the experimental results, as pointed out by (Hull, 1993). Nevertheless, current methodologies do not imply any particular coordination or synchronization between the basic scientific data and the analyses on them, which are treated as almost separate items. On the contrary, researchers would greatly benefit from an integrated vision of them provided by means of a DLS, where the access to a scientific data item could also offer the possibility of retrieving all the analyses and interpretations on it. TrebleCLEF will not only take care of managing a scientific DLS as has been done in recent CLEF campaigns but also of providing the tools and the methodologies for further processing and comparing the collected data through statistical analyses, such as those adopted in (Di Nunzio et al, 2006), as well as by adopting methods specifically developed for the multilingual context (Crivellari, Di Nunzio & Ferro, 2007).

4.2 User-oriented Studies

User studies have received less attention from the scientific community, partly due to their cost (much higher than running batch experiments) and partly due to the difficulties of establishing evaluation methodologies which are both realistic (performed in real-world scenarios) and scientifically well-grounded (performed under laboratory-controlled conditions) (Petrelli, 2007). TrebleCLEF will address the needs of (at least) three types of users with strong interests in an evaluation forum and its end products: a) multilingual system developers; b) business companies with a potential interest in MLIA system software (the potential market for system developers); c) end users with information needs that transcend language barriers

An important part of the user studies will be query log analysis. The logs of operational search engines will be analysed to study users’ patterns of search. A valuable source of this information will be the logs supplied by The European Library, a service which offers access to the resources of 47 national libraries of Europe in twenty different languages [6]. Such analyses allow lab-based testing of the systems of large search engines involving the interactions of millions of users: a scale of user evaluation inconceivable to previous user study research. Logs for a particular search engine can be analysed, patterns of use determined, then changes can be made to the engine and patterns of use re-examined to determine the impact of the change. In addition, large sets of logs can be split into training and testing sets. User models can be built from examination of user interaction in the training set and the models can be used to predict how users will search in the test set.

4.3 Test Collection Creation

The common approach to lab-based evaluation is through use of test collections. Europe has often been a leader in this research field, with one of the first large scale evaluation exercises occurring at Cranfield (UK) in the late 1960s (Cleverdon, 1967); drawing in researchers from all over the world. European researchers then led the way in describing how evaluation should move beyond the Cranfield experiments to much larger collections (Spärck Jones & Van Rijsbergen, 1975) This research defined the structure for the well known US-based TREC evaluations, which spawned a number of spin off exercises, including CLEF. Although such large collaborative evaluation exercises still have great value in defining the worth of broad-based approaches to improving search technology, it is clear from personal communications from those working in large search engine companies, that careful individual testing of search technologies is critical to building successful search systems. It is even the case when a top performing search engine is moved from one collection (say documents written in one language) to another (documents written in a different language) that the engine

[6] www.theeuropeanlibrary.org

needs to be re-configured using a process of careful evaluation.

It is generally assumed by many researchers that constructing test collections demands great effort and can only be afforded by rich organisations or through extensive collaboration with large numbers of researchers. Current evaluation campaigns reinforce this belief. However, such attitudes ignore the flood of research currently being conducted on new measures and new methodologies that allow building test collections more efficiently (Sanderson & Joho, 2004; Buettcher, et al, 2007) along with new measures that work well with the new test collections. TrebleCLEF aims at identifying and collating the latest research in methods for forming test collections quickly and efficiently and at identifying new evaluation methodologies and metrics specifically designed and tuned for use in a multilingual context.

4.4 Grid-Experiments

The impact of language-dependent issues on the performance of information retrieval systems is still not fully understood. Individual researchers or small groups do not usually have the possibility of running large-scale and systematic experiments over a large set of experimental collections and resources in order to improve the comprehension of MLIA systems and gain an exhaustive picture of their behaviour with respect to languages. TrebleCLEF will address this lack of information by promoting and coordinating a series of systematic so-called “grid-experiments” which will re-use and exploit the valuable resources and experimental collections made available by CLEF in order to gain more insights about the effectiveness of the various weighting schemes and retrieval techniques with respect to the languages and to disseminate this knowledge to the relevant application communities.

4.5 Language Resources for MLIA

TrebleCLEF will support the development of high priority language resources[7] for Multilingual Information Access in a systematic, standards-driven, collaborative learning context. Priority requirements will be assessed through consultations with language industry and communication players, and a protocol and roadmap will be established for developing a set of language resources for all technologies related to MLIA.

5. Conclusions

To sum up, TrebleCLEF will not only sponsor R&D and evaluation in the multilingual retrieval context but will focus on those aspects of system implementation that have been somewhat neglected with the aim of preparing an exhaustive set of best practice recommendations addressing the issues involved from both the system and the user perspective. The goal is to disseminate the research findings to system developers encouraging easy take up of MLIA technology by the application communities.

6. References

Agosti, M., Di Nunzio, G. M., Ferro, N., Harman, D., & Peters, C. (2007). The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe. In *Proceedings ECDL 2007).* LNCS 4675, Springer, Heidelberg, Germany, pp 509–512.

Braschler, Martin (2004). Robust Multilingual Information Retrieval, Dissertation, Université de Neuchatel.

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

Buettcher, S., Clarke C., Yeung P., Soboroff I. (2007) Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements, in the *Proceedings of ACM SIGIR Conference*

Cleverdon, C.W. (1967) The Cranfield tests on index language devices. *Aslib Proceedings*, 19, pp 173-192

Crivellari, F., Di Nunzio, G. M., and Ferro, N. (2007). How to Compare Bilingual to Monolingual Cross-Language Information Retrieval. In Amati, G., Carpineto, C., and Romano, G., Eds, *In Proceedings ECIR 2007*, LNCS 4425, Springer, Heidelberg, Germany, pp 533–540

Di Nunzio, G. M., & Ferro, N. (2006). Scientific Evaluation of a DLMS: a service for evaluating information access components. In J. Gonzalo, C. Thanos, M. F. Verdejo, & R. C. Carrasco (Eds.), *Proc. ECDL 2006*. LNCS 4172, Springer, Heidelberg, Germany, pp 536–539.

Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2006). CLEF 2006: Ad Hoc Track Overview. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, et al. (Eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval : Seventh Workshop of the Cross--Language Evaluation Forum (CLEF 2006).*. LNCS 4730, Springer, Heidelberg, Germany, pp 21–34.

Dussin, M., & Ferro, N. (2007). Design of the User Interface of a Scientific Digital Library System for Large-Scale Evaluation Campaigns. In Thanos, C. & Borri, F. (Eds.), *Working Notes of the Second DELOS Conference*.

Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In Korfhage, R., Rasmussen, E., and Willett, P., editors, *Proc. SIGIR 1993*, pp 329–338. ACM Press, New York, USA.

Petrelli, D. (2008) On the role of User-Centred Evaluation in the Advancement of Interactive Information Retrieval. *Information Processing and Management*, 44(1), pp 22-38.

Sanderson, M., Joho, H. (2004) Forming test collections with no system pooling, in the *27th ACM SIGIR Conference*, pp 33-40

Spärck Jones, K., van Rijsbergen, C.J. (1975) Report on the need for and provision of an “ideal” information retrieval test collection, *British Library Research and Development Report*

[7] Language resources have to be interpreted in a wider context: text data, speech data but also modalities and media beyond the language such as video, acoustics, images, etc.