The Future of Large-Scale Evaluation Campaigns for Information Retrieval in Europe

Maristella Agosti[1], Giorgio Maria Di Nunzio[1], Nicola Ferro[1], Donna Harman[2], and Carol Peters[3]

[1] Department of Information Engineering, University of Padua, Italy
{agosti, dinunzio, ferro}@dei.unipd.it
[2] National Institute of Standards and Technology, USA
donna.harman@nist.gov
[3] ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract. A Workshop on "The Future of Large-scale Evaluation Campaigns" was organised jointly by the University of Padua and the DELOS Network of Excellence and held in Padua, Italy, March 2007. The aim was to perform a critical assessment of the scientific results of such initiatives and to formulate recommendations for the future. This poster summarises the outcome of the discussion with respect to the major European activity in this area: the Cross Language Evaluation Forum.

1 Motivations for the Workshop

Since its beginnings, the DELOS Network of Excellence[1] has recognised the importance of R&D in the area of multilingual information access (MLIA) and has supported the activities of the Cross Language Evaluation Forum (CLEF).

CLEF began life as a track for cross-language system evaluation within the Text REtrieval Conference (TREC) series[2] but was launched as a separate European activity in 2000 with the goal of promoting the development of MLIA functionality, producing test collections for system benchmarking, and last but not least creating a multidisciplinary research community around this domain.

A major aim of a workshop held in Padua in March 2007 was to assess the achievements of CLEF over the years, and to discuss directions for an eventual continuation under FP7. Commonalities and differences between CLEF and TREC were also examined. This poster aims at stimulating further discussion and getting feedback on the future of CLEF.

[1] DELOS is currently running under the European Commission's Sixth Framework programme (FP6), see http://www.delos.info/
[2] TREC is the major initiative for information retrieval system evaluation in North America, see http://trec.nist.gov/

L. Kovács, N. Fuhr, and C. Meghini (Eds.): ECDL 2007, LNCS 4675, pp. 509–512, 2007.

2 Achievements of CLEF

When CLEF first started, the few existing cross-language information systems generally handled only two languages, one of which was normally English, and ran only for textual document retrieval. The long-term goal of CLEF has thus been to promote the development of truly multilingual, multimodal systems via a systematic study of the requirements of digital libraries and other globally distributed information repositories, and the design of tasks that meet these needs. Over the years, we have gradually introduced new tracks and more complex tasks to assess free-text and domain-specific cross-language retrieval, multiple language question answering, cross-language retrieval for speech and for image collections, multilingual retrieval of web documents, and cross-language geographic retrieval. For complete details of the CLEF 2000 - 2007 agendas, see the website at `http://www.clef-campaign.org/`.

From discussions at the Padua workshop, it was established that the main achievements of CLEF over the years can be summarised in the following points:

- implementation of a powerful and flexible technical infrastructure including data curation functionality;
- promotion of research in previously unexplored areas, such as cross-language question answering, image and geographic information retrieval;
- improvement in performance for cross-language text retrieval systems (from 50% of monolingual retrieval in 2000 to at least 85% in 2006);
- quantitative and qualitative evidence with respect to user interaction and best practice in cross-language system development;
- creation of important, reusable test collections for system benchmarking, covering 12 languages and three media (text, speech and image);
- building of a strong, multidisciplinary research community (94 groups from 5 continents submitted results in 2006).

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. For a more detailed assessment of CLEF, see [1].

It was agreed that CLEF has been crucial in stimulating research in multilingual IR not only in Europe, impacting both the information retrieval and the digital library research areas.

3 Recommendations for the Future

However, despite these achievements, it was recognised by the participants at the workshop that future editions of CLEF should not only continue to support annual system evaluation campaigns with tracks and tasks designed to stimulate R&D in the MLIA domain but should also (i) develop the facilities to further exploit the results of these campaigns by promoting in-depth studies and analyses of the outcomes, (ii) focus on areas of research previously ignored by CLEF mainly due to lack of resources, (iii) encourage the dissemination and technology transfer of the

results obtained to the European digital library and related communities through the specification of best practices in MLIA system development.

With respect to the first point, it is recognised that 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. 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. Researchers would greatly benefit from an integrated vision of data plus analyses provided by means of a scientific digital library system, where access to a scientific data item could also offer the possibility of retrieving all the analyses and interpretations on it [2].

As CLEF is run almost entirely as a voluntary exercise it is not always easy to find the necessary resources to follow a given line of action. However, if possible, we believe that efforts in the near future should be concentrated in the following key research areas:

- user modelling, 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 component development 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.

We also need to identify new metrics specifically designed and tuned for use in a multilingual context and we need to study new methods for creating test collections quickly and efficiently [3]. So far, most CLEF evaluation methodologies have tended to adapt and reuse evaluation methodologies already experimented at TREC. We must move beyond topic-based relevance, absolute relevance of documents in isolation and mean average precision to include multi-valued criteria, such as diversity, novelty, authority, recency, and to address tasks which still do not have well-developed evaluation methodologies. In particular, we need to work on establishing realistic and scientifically well-grounded evaluation methodologies for interactive MLIA experiments and user studies [4].

In fact, a criticism made of CLEF at the workshop is that so far we have focussed too much on measuring overall system performance according to ranked lists of results while neglecting many other important aspects. As mentioned above, one area that needs to be addressed in far greater depth is that of user-centred evaluation; we need to know whether the system performance actually satisfies the user expectations? For this reason, we believe that, in the future the interactive track should be extended and more attention given to aspects involving user satisfaction issues. One question is whether average precision is really the best metric from the user viewpoint. In CLEF 2007, new metrics have been introduced into the ad-hoc track in order to favour systems that achieve a high precision of correct responses in the first ten results returned - rather than a good average precision. This is a user-oriented measure and we believe makes more sense in the Internet dominated world.

Another issue regards system response times. CLEF 2006 took a first step in this direction with the organization of a real-time exercise as part of the question-answering track. In the end, the question was whether the best multilingual question answering system was the fastest system or the most accurate one and, given the choice, would the user prefer a faster system over a slightly less accurate but slower one.

An important point made at the workshop was that there is still very little take-up of MLIA functionality by the market. In fact, 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. One of the challenges that CLEF must face in the near future is how to best transfer the research results to the market place. In our opinion, if the gap between academic excellence and commercial adoption of MLIA technology is to be bridged, we need to extend the current CLEF formula in order to give application communities the possibility to benefit from the CLEF evaluation infrastructure without the need to participate in academic exercises that may be irrelevant to their current needs. We feel that CLEF should introduce an application support structure aimed at encouraging take-up of the technologies tested and optimized within the context of the evaluation exercises. This structure would provide tools, resources, best practice guidelines and consulting services to applications or industries that need to include multilingual functionality within a service or product.

In summary, CLEF should function as a center of competence for European multilingual information retrieval system research, development, implementation, and related activities.

Acknowledgements

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, in the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

1. Agosti, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF: Ongoing Activities and Plans for the Future. In: Proc. 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NII, Tokyo, Japan, pp. 493–504 (2007)
2. Agosti, M., Di Nunzio, G.M., Ferro, N.: A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In: Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), NII, Tokyo, Japan, pp. 62–73 (2007)
3. Sanderson, M., Joho, H.: Forming Test Collections with No System Pooling. In: Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pp. 33–40. ACM Press, New York (2004)
4. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer, Heidelberg (2005)