A PROMISE for Experimental Evaluation

Martin Braschler[1], Khalid Choukri[2], Nicola Ferro[3], Allan Hanbury[4], Jussi Karlgren[5], Henning Müller[6], Vivien Petras[7], Emanuele Pianta[8], Maarten de Rijke[9], and Giuseppe Santucci[10]

[1] Zurich University of Applied Sciences, Switzerland
martin.braschler@zhaw.ch
[2] Evaluations and Language resources Distribution Agency, France
choukri@elda.org
[3] University of Padua, Italy
ferro@dei.unipd.it
[4] Information Retrieval Facility, Austria
a.hanbury@ir-facility.org
[5] Swedish Institute of Computer Science, Sweden
jussi@sics.se
[6] University of Applied Sciences Western Switzerland, Switzerland
henning.mueller@sim.hcuge.ch
[7] Humboldt-Universität zu Berlin, Germany
vivien.petras@ibi.hu-berlin.de
[8] Centre for the Evaluation of Language Communication Technologies, Italy
pianta@fbk.eu
[9] University of Amsterdam, The Netherlands
derijke@uva.nl
[10] Sapienza University of Rome, Italy
santucci@dis.uniroma1.it

Abstract. *Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE)* is a Network of Excellence, starting in conjunction with this first independent CLEF 2010 conference, and designed to support and develop the evaluation of multilingual and multimedia information access systems, largely through the activities taking place in *Cross-Language Evaluation Forum (CLEF)* today, and taking it forward in important new ways.

PROMISE is coordinated by the University of Padua, and comprises 10 partners: the Swedish Institute for Computer Science, the University of Amsterdam, Sapienza University of Rome, University of Applied Sciences of Western Switzerland, the Information Retrieval Facility, the Zurich University of Applied Sciences, the Humboldt University of Berlin, the Evaluation and Language Resources Distribution Agency, and the Centre for the Evaluation of Language Communication Technologies.

The single most important step forward for multilingual and multimedia information access which PROMISE will work towards is to provide an *open evaluation infrastructure* in order to support *automation* and *collaboration* in the evaluation process.

1 Multilingual and Multimedia Information Access

With a population of over 500 million in its 27 states, EU citizens and companies demand information access systems that allow them to interact with the culturally and politically diverse content that surrounds them in multiple media. Currently, about 10 million Europeans work in other member states of the Union, and alongside the Union's 23 official languages and 3 alphabets, 60 additional regional or group languages are used, as well as approximately one hundred languages brought by immigrants [2]. Most of the Union's inhabitants know more than a single language, and the stated political aim is for every citizen to be able to use their first language and two additional languages in their professional and everyday tasks.

With the advance of broadband access and the evolution of both wired and wireless connectivity, today's users are not only information consumers, but also information *producers*: they create their own content, augment existing material through annotations (e.g. adding tags and comments) and links, and mix and mash up different media and applications within a dynamic and collaborative information space. The expectations and habits of users are constantly changing, together with the ways in which they interact with content and services, often creating new and original ways of exploiting them.

In this evolving scenario, language and media barriers are no longer necessarily seen as insurmountable obstacles: they are constantly being crossed and mixed to provide content that can be accessed on a global scale within a multicultural and multilingual setting. Users need to be able to co-operate and communicate in a way that crosses language and media boundaries and goes beyond separate search in diverse media and languages, but that exploits the interactions between languages and media.

To build the tools of the future that we already see taking shape around us, experimental evaluation has been and will continue to be a key means for supporting and fostering the development of multilingual and multimedia information systems. However this evaluation cannot solely be based on laboratory benchmarking, but needs to involve aspects of usage to *validate* the basis on which systems are being designed, and needs to involve stakeholders beyond the technology producers: media producers, purveyors, and consumers.

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE)[1] aims at advancing the experimental evaluation of complex multimedia and multilingual information systems in order to support individuals, commercial entities, and communities who design, develop, employ and improve such complex systems. The overall goal of PROMISE is to deliver a unified and open environment bringing together data, knowledge, tools, methodologies, and development and research communities, as measured by three criteria:

[1] `http://www.promise-noe.eu/`

1. Increase in the volume of evaluation data;
2. Increase in the rate of the utilization of the data by development and research sites;
3. Decrease in the amount of the effort needed for carrying out evaluation and tests.

As a partial consequence of the lowered effort, we expect to find larger uptake and activities in each task. We expect to find a *larger community* of usage, with more sites participating in various scales of experimental participation.

As a third and most important long-term effect we confidently expect to see new and sustainable results emerge from the evaluation process, being deployed fruitfully in research and engineering efforts across the world.

2 Open Evaluation Infrastructure

Evaluation, while often done using high-end computational tools, has mostly been a manual process. This has been a bottleneck for scaling up in volume, but most of all in porting and sharing methodologies and tools. PROMISE aims to help, by providing an infrastructure for the effective automation of evaluation. PROMISE will develop and provide an open evaluation infrastructure for carrying out experimentation: it will support all the various steps involved in an evaluation activity; it will be used by the participants in the evaluation activities and by the targeted researcher and developer communities; it will collect experimental collections, experimental data, and their analyses in order to progressively create a knowledge-base that can be used to compare and assess new systems and techniques. This work will be partially based on previous efforts such as the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* system [1]. The proposed open evaluation infrastructure will make topic creation, creation of pools and relevance assessment more efficient and effective. PROMISE expects noticeable scale increases in each of these aspects. Moreover, it will provide the means to support increased collaboration among all the involved stakeholders, e.g. by allowing them to annotate and enrich the managed contents, as well as to apply information visualization techniques [4] for improving the representation and communication of the experimental results.

3 Use Cases as a Bridge between Benchmarking and Validation

Benchmarking, mostly using test collections such as those provided by CLEF, has been a key instrument in the development of information access systems. Benchmarking provides a systematic view of differences between systems, but only if the basic premises of usage can be held constant. Moving from the task- and topic-based usage for which past information systems have been developed to multilingual and multimedia systems for a diverse user population, these premises are being contested. To better serve the development of future systems,

future evaluation efforts need to be explicit about what aspects of usage the systems are expected to provide for, and evaluation needs to be tailored to meet these requirements. To move from abstract benchmarking to more user-sensitive evaluation schemes, PROMISE will, in its requirement analysis phase, formulate a set of *use cases* based on usage scenarios for multimedia and multilingual information access. This will allow future development and research efforts to identify similarities and differences between their project and any predecessors, and address these through choices not only with respect to technology but also with respect to projected usage. This will mean a more elaborate requirements analysis and some care in formulating target notions, evaluation metrics, and comparison statistics.

PROMISE will begin by addressing three initial use cases, and more will be systematically added during the project:

Unlocking culture will deal with information access to cultural heritage material held in large-scale digital libraries, based on the *The European Library (TEL)* Collection: a multilingual collection of about 3 million catalogue records.

Search for innovation will deal with patent search and its requirements, based on MAREC (MAtrixware REsearch Collection): a collection of 19 million patent applications and granted patents in 5 languages.

Visual clinical decision support will deal with visual information connected with text in the radiology domain, based on the ImageCLEF collection with currently 84 000 medical images.

4 Continuous Experimentation

Today, evaluation in organised campaigns is mainly unidirectional and follows annual or less frequent cycles: system owners have to download experimental collections, upload their results, and wait for the performance measurements from the organisers. To overcome this, the open architecture provided by PROMISE will allow for two different modes of operation:

1. The open evaluation infrastructure can be remotely accessed: system owners will be able to operate continuously through standard interfaces, run tests, obtain performance indicators, and compare them to existing knowledge bases and state-of-the-art, without necessarily having to share their own results;
2. In the case of a system implementing a set of standard interfaces, the infrastructure will be able to directly operate the system and run a set of tests to assess its performances, thus speeding up the adoption of standard benchmarking practices.

In this way, PROMISE will automate the evaluation process and transform it from periodic to continuous, radically increasing the number of experiments that can be conducted and making large-scale experimental evaluation part of the

daily tools used by researchers and developers for designing and implementing their multilingual and multimedia information systems. The open architecture will also enable commercial developers to use the same evaluation methodologies as research projects do, without necessitating public disclosure of results.

5 PROMISE and CLEF

CLEF is a renowned evaluation framework, which has been running for a decade, with the support of major players in Europe and, in general, in the world. CLEF 2010 represents a renewal of the "classic CLEF" format and an experiment to understand how "next generation" evaluation campaigns might be structured [3]. PROMISE will continue the innovation path initiated for CLEF, will provide infrastructure for CLEF tasks, and will extend the scope of CLEF, through an open evaluation infrastructure; through use-case driven experimentation; through an automated, distributed, and continuous evaluation process and through tools for collaboration, communication, sharing, leading to further community building between the research groups that have turned CLEF into the success that it is.

References

1. M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In G. Tsakonas and C. Papatheodorou, editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK, 2009.
2. Commission of the European Communities. Multilingualism: an asset for Europe and a shared commitment. *COMM(2008) 566 Final*, September 2008.
3. N. Ferro. CLEF, CLEF 2010, and PROMISEs: Perspectives for the Cross-Language Evaluation Forum. In N. Kando and K. Kishida, editors, *Proc. 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 2–12. National Institute of Informatics, Tokyo, Japan, 2010.
4. D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in Visual Data Analysis. In E. Banissi, editor, *Proc. of the 10th International Conference on Information Visualization (IV 2006)*, pages 9–16. IEEE Computer Society, Los Alamitos, CA, USA, 2006.