CLEF 2010 Labs

Call for Participation

22–23 September 2010, Padua, Italy http://www.clef2010.org/

CLEF 2010

Is the continuation of the popular CLEF campaigns that have run for the past ten years. It will consist of two main parts: a **peer-reviewed conference** on multilingual and multimodal information access evaluation and a series of **labs** which will continue the CLEF tradition of community-based evaluation and discussion on evaluation issues.

Workshop

The labs will culminate in sessions of a half-day, one full day or two days workshop at the CLEF 2010 conference.

Publication

The Working Notes of the LABs will be published online in time for the conference. It is foreseen that this online publication will have an ISBN number and be indexed in relevant services.

Participation

Registration to the labs is via the CLEF website http://www.clef2010.org. Registration remains open until 15 May 2010.

Coordination, Contact Information

Lab Chairs

Martin Braschler, ZHAW, Switzerland
Donna Harman, NIST, USA.

Organization Chair

Emanuele Pianta, CELCT, Italy.

Resource Chair

Khalid Choukri, ELDA, France.

Two different forms of labs are offered: *benchmarking activities* which are very similar to the CLEF "tracks" evaluation campaigns, and *workshop-style* labs that explore issues of information access evaluation and related fields.

Benchmarking activities

CLEF-IP

This year, CLEF-IP puts to use a collection of almost 2 million patent documents in XML format with content in English, German, and French.
The lab offers a Prior Art Candidate Search task and a Classification task.
The first task will ask participants to retrieve documents that are potential prior art to a given document. Topics will be chosen as to stimulate multilingual retrieval. The second task will ask participants to classify documents according to the International Patent Classification scheme. Training data for both tasks will be available prior to the topic sets release. Relevance assessment will be done using patent citations for the first task and current patent classifications for the second task.
Coordinator: Information Retrieval Facility (AT). See http://www.ir-facility.org/research/evaluation/clef-ip-10

Cross-Language Image Retrieval (ImageCLEF)

This track evaluates retrieval from visual collections; both text and visual retrieval techniques are exploitable. Four challenging tasks are foreseen: 1) retrieval from a Wikipedia collection containing images and structured information in several languages; 2) medical image retrieval with visual, semantic and mixed topics in several languages with a data collection from the scientific literature; 3) detection of semantic categories from robotic images (non-annotated collection, concepts to be detected); 4) a photo annotation task that investigates automated semantic annotation based on visual information with approaches based on Flickr user tags and multimodal approaches. Track coordinators are U. of Applied Sciences Western Switzerland (CH), Oregon Health and Science U. (US), CWI (NL), TELECOM Bretagne (FR), Leiden University (NL), U. of Geneva (CH), Fraunhofer Society (DE), IDIAP (CH). See also http://www.imageclef.org/.

Uncovering Plagiarism, Authorship, and Wikipedia Vandalism (PAN)
New this year

PAN @ CLEF 2010 divides into two tasks:
- Plagiarism Detection. Today's plagiarism detection systems are faced with intricate situations, such as obfuscated plagiarism or plagiarism within and across languages. Moreover, the source of a plagiarism case may be hidden in a large collection of documents, or it may not be available at all. Following the success of the 2009 campaign on plagiarism detection, we will provide a revised evaluation corpus consisting of artificial and simulated plagiarism.
- Wikipedia Vandalism Detection. Vandalism has always been one of Wikipedia's biggest problems. However, the detection of vandalism is done mostly manually by volunteers, and research on automatic vandalism detection is still in its infancy. Hence, solutions are to be developed which aid Wikipedians in their efforts.

The lab is organized by the Bauhaus-Universität Weimar, the Universidad Politécnica de Valencia, the University of the Aegean, and the Bar-Ilan University. See http://pan.webis.de for more details.

CLEF 2010 Labs

Call for Participation

22–23 September 2010, Padua, Italy http://www.clef2010.org/

Benchmarking activities

ResPubliQA

Two separate tasks are proposed for the ResPubliQA 2010 evaluation campaign which allow both passages and exact answers (smallest exact demarcation) to be returned as the type of answer in response to the same 200 input questions. Systems can also return NOA if they are not confident of their answer.
The focus is on the direct comparison of systems' performances among languages, a goal which is enabled by the adoption of the multilingual parallel paragraph-aligned document collections (JRC-Acquis and Europarl) of EU legislative documents, available in 9 languages, (i.e: Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish).
The Lab is jointly coordinated by UNED, CELCT and the University of Limerick.
For more details visit the website http://celct.isti.cnr.it/ResPubliQA/

WePS - *New this year*

WePS-3 is a competitive evaluation campaign which consists of two tasks concerning the Web Entity Search Problem:
- Task 1 is related to *Web People Search*, and focuses on person name ambiguity and person attribute extraction on Web pages. Given a set of web search results for a person name, the tasks consists of clustering the pages according to the different people sharing the name and extract certain biographical attributes for each person.
- Task 2 is related to *Online Reputation Management* for organizations, and focuses on the problem of ambiguity for organization names and the relevance of Web data for reputation management purposes. Given a set of Twitter entries containing an (ambiguous) company name, and given the home page of the company, the tasks consists of discriminating entries that do not refer to the company.

WePS 3 is coordinated by three universities (UNED, New York University, the University of Illinois at Chicago) and two corporate stakeholders: Intelius Corp. and Llorente & Cuenca. More information at http://nlp.uned.es/weps.

Workshops

Cross-lingual Expert Search - Bridging CLIR and Social Media (CriES) - *New this year*

This workshop addresses the problem of multi-lingual expert search in social media environments. The main topics are multi-lingual expert retrieval methods, social media analysis with respect to expert search, selection of datasets and evaluation of expert search results.
In addition to the workshop we also organize a pilot challenge:
- *Workshop*: We expect submissions addressing the main topics including user characterization in multi-lingual social media, community analysis for retrieval scenarios, user-centric recommender algorithms, proposals of new social media datasets and evaluation of cross-lingual expert search.
- *Pilot Challenge*: The challenge is based on a dataset from Yahoo!Answers, consisting of multi-lingual questions, answers and user relations. Given a set of multi-lingual questions the task is to retrieve relevant users that will most likely be able to answer the questions.

Coordinators are KIT, U. of Koblenz and U. of Bielefeld (DE). See http://www.multipla-project.org/cries for details.

LogCLEF

The goal of LogCLEF is the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems.
A common data set will be distributed to the participants. In coordination with the organizers, participating groups will be devoted to different tasks in exploring and understanding the data. Tasks will include the identification of the language of a query, identification of sessions with more than one language, user clustering, labelling named entities (esp. person names and geographic names) and linking them to these entities (e.g. Wikipedia pages). Both search log and HTTP logs for the 2007/2008 (the same period used this year), plus the search log of 2009 will be (most likely) available from The European Library.
At the workshop, participants are required to present their algorithms, their results and discuss what the results tell about user behavior. The workshop will be the basis for a definition of a set of competitive tasks for future studies on log analysis at LogCLEF.
Coordinators are: University of Hildesheim and University of Padua. See http://www.uni-hildesheim.de/logclef/