QA4MRE 2012: Question Answering for Machine Reading Evaluation

An Evaluation Lab at CLEF 2012

http://celct.fbk.eu/QA4MRE/

First Call for Participation

Building on the experience of last year campaign, we invite teams to participate to QA4MRE at CLEF 2012[1], an evaluation campaign of Machine Reading systems through Question Answering and Reading Comprehension Tests. Systems should be able to use knowledge from given texts as a way to answer a set of questions. The task focuses on the reading of single documents and the identification of the answers to a set of questions about information that is stated or implied in the text. Questions are in the form of multiple choice, each having five options, and only one correct answer. The participating systems will be required to answer the questions by choosing in each case one answer from the five alternatives.
The detection of correct answers is specifically designed to require various kinds of inference and the consideration of previously acquired background knowledge from reference document collections provided by the organization. Although the additional knowledge obtained through the background collection may be used to assist with answering the questions, the principal answer is to be found among the facts contained in the test documents given.

Languages
Background collections, test documents and reading tests will be available in Arabic, Bulgarian, English, German, Italian, Romanian, and Spanish.

- The background collections created in the different languages will be available to all participants, signing a license agreement. Thus, the learning of knowledge could be in one language or several.
- The tests will be exactly the same in all languages, using parallel translations.

This year, in addition to the main task, two pilot tasks will be offered, namely:

1. *Processing Modality and Negation for Machine Reading:* aimed at evaluating whether systems are able to understand extra-propositional aspects of meaning like modality and negation. Modality is a grammatical category that allows expressing aspects related to the attitude of the speaker towards his/her statements. Modality understood in a broader sense is also related to the expression of certainty, factuality, and evidentiality. Negation is a grammatical category that allows changing the truth value of a proposition. Our plan is to integrate modality and negation in the main task next year. More info at: http://celct.fbk.eu/QA4MRE/index.php?page=Pages/modalityTask.html

2. *Machine Reading of Biomedical Texts about Alzheimer*: aimed at setting questions in the Biomedical domain with a special focus on one disease, namely Alzheimer. This pilot task will explore the ability of a system to answer questions using scientific language. Texts will be taken from PubMed Central related to Alzheimer and from Medline abstracts. PMC is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM). MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. It was compiled by the United States National Library of Medicine (NLM), and is freely available on the Internet. Additionally, the background collection will contain also a collection of articles about the key hypotheses in Alzheimer Disease published by Elsevier.
   In order to keep the task reasonably simple for systems, participants will be given the background collection already processed with Tok, Lem, POS, NER, and Dependency parsing. A development set will

[1] http://clef2012.org/

also be provided to participants. More info at :
http://celct.fbk.eu/QA4MRE/index.php?page=Pages/biomedicalTask.html

Both pilot tasks will be offered in **English** only and will be coordinated by the University of Antwerp, Belgium.

The results of the evaluation campaign will be disseminated at the final workshop which will be organized in conjunction with CLEF 2012, 17–20 September in Rome, Italy.

Evaluation

Evaluation will be performed automatically by comparing the answers given by systems to the ones given by humans. Each test will receive an evaluation score between 0 and 1 using c@1. This measure, already tried in previous CLEF QA Tracks, encourages systems to reduce the number of incorrect answers and maintain the number of correct ones by leaving some questions unanswered.

Tentative Schedule

- Track Guidelines: **January, 31**
- Release of background collections in all the languages of the task: **April, 6**
- Test set release: **June, 5**
- Run submissions: **June, 15**
- Individual Results to Participants: **June, 25**
- Working Notes Papers: according to CLEF schedule

General Coordination

- Anselmo Peñas, UNED NLP & IR Group, Spain
- Eduard Hovy, Information Sciences Institute, USA

Technical Coordination

- Pamela Forner, CELCT, Italy
- Alvaro Rodrigo, UNED NLP & IR Group, Spain
- Richard Sutcliffe, University of Limerick, Ireland

Pilot Tasks Coordination

- Roser Morante, University of Antwerp
- Walter Daelemans, University of Antwerp

Organizing Committee

- Yassine Benajiba, Philips Research North America
- Corina Forascu, Alexandru Ioan Cuza University, Romania
- Petya Osenova, Bulgarian Academy of Sciences, Bulgaria
- Caroline Sporleder, University of Saarland, Germany

Programme Committee

- Ken Barker, University of Texas at Austin, USA
- Johan Bos, Rijksuniversiteit Groningen, Netherlands
- Peter Clark, Vulcan Inc., USA
- Ido Dagan, Bar-Ilan University
- Bernardo Magnini, Fondazione Bruno Kessler, Italy
- Dan Moldovan, University of Texas at Dallas, USA
- Emanuele Pianta, Fondazione Bruno Kessler, and CELCT, Italy
- John Prager, IBM, USA
- Hoa Trang Dang, NIST, USA
- Dan Tufis, Research Institute for Artificial Intelligence, Romanian Academy