QA4MRE 2013: Question Answering for Machine Reading Evaluation

An Evaluation Lab at CLEF 2013

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

First Call for Participation

Building on the experience of last year's campaign, we invite teams to participate to QA4MRE at CLEF 2013[1], an evaluation campaign of Machine Reading systems through Question Answering and Reading Comprehension Tests. Systems should be able to use knowledge obtained automatically from given texts to answer a set of questions posed for single documents at a time. Questions are in the form of multiple choice, where a significant portion of questions have no correct answer among the given alternatives proposed. While the principal answer is to be found among the facts contained in the test documents provided, systems may use knowledge from additional given texts (the 'Background Corpus') to assist them with answering the questions. Some questions will also test a system's ability to understand certain propositional aspects of meaning such as modality and negation.

Languages
Test documents and reading tests will be available in Arabic, Bulgarian, English, Romanian, and Spanish.
- The tests will be exactly the same in all languages, using parallel translations.

Background Collections
- The background collections of additional text about the domain, created in the different languages, will be available to all participants who sign a license agreement. Thus, the learning and use of additional knowledge could be in one language or several.
- The 2013 background collections are based on but not identical to the 2012 collections.

Three tasks will be offered, namely:

1. *QA4MRE main task*: aimed at answering a series of multiple choice tests, each based on a single document, about four general language topics

2. *Machine Reading of Biomedical Texts about Alzheimer's Disease*: aimed at setting questions in the Biomedical domain with a special focus on one disease, namely Alzheimer. This 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 also be provided to participants. More info at:

---

[1] http://clef2013.org/

The task will be offered in **English** only and will be coordinated by the University of Antwerp, Belgium.

3. *Entrance Exams*

Japanese University Entrance Exams include questions formulated at various levels of complexity and test a wide range of capabilities. The challenge of "Entrance Exams" aims at evaluating systems under the same conditions humans are evaluated to enter the University. In this first campaign we will reduce the challenge to Reading Comprehension exercises contained in the English exams. More types of exercises will be included in subsequent campaigns (2014–2016) in coordination with the "Entrance Exams" task at NTCIR. Exams are created by the Japanese National Center for University Admissions Tests. The "Entrance Exams" corpus is provided by NII's Todai Robot Project and NTCIR.

The results of the evaluation campaign will be disseminated at the final workshop which will be organized in conjunction with CLEF 2013, 23–26 September in Valencia, Spain.

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, 30**
- Release of background collections in all the languages of the task: **March, 8**
- Test set release: **May, 6**
- Run submissions: **May, 15**
- Individual Results to Participants: **May, 20**
- Working Notes Papers: **June, 15**

General Coordination
- Anselmo Peñas, UNED NLP & IR Group, Spain
- Eduard Hovy, Language Technologies Institute, Carnegie Mellon University, Pittsburgh USA

Technical Coordination (Main Task)
- Pamela Forner, CELCT, Trento, Italy
- Alvaro Rodrigo, UNED NLP & IR Group, Madrid, Spain
- Richard Sutcliffe, School of CSEE University of Essex Colchester, United Kingdom

Machine Reading of Biomedical Texts about Alzheimer's Disease
- Roser Morante and Walter Daelemans, University of Antwerp, Belgium

Entrance Exams
- Anselmo Peñas, UNED, Spain
- Eduard Hovy, Carnegie Mellon University, USA
- Pamela Forner, CELCT, Italy
- Noriko Kando, National Institute of Informatics, Japan
- Teruko Mitamura, Carnegie Mellon University, USA
- Yusuke Miyao, National Institute of Informatics, Japan

Organizing Committee (Main Task)

- Yassine Benajiba, Thomson Reuters, USA
- Corina Forascu, Alexandru Ioan Cuza University, Romania
- Petya Osenova, Bulgarian Academy of Sciences, Bulgaria

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, Israel
- Bernardo Magnini, Fondazione Bruno Kessler, Italy
- Dan Moldovan, University of Texas at Dallas, USA
- John Prager, IBM, USA
- Hoa Trang Dang, NIST, USA
- Dan Tufis, Research Institute for Artificial Intelligence, Romanian Academy, Romania