QA4MRE
Question Answering for Machine Reading Evaluation
at CLEF 2011

Track Guidelines

INTRODUCTION / MOTIVATION

Machine Reading can be defined as the automatic understanding of text[1]. One way of evaluating the understanding of a text is to assess a system's ability to answer a set of questions about it.

The methodology proposed in this evaluation campaign represents an evolution of previous evaluation approaches in Natural Language Processing (NLP), i.e. Question Answering[2], Recognizing Textual Entailment[3], and Answer Validation[4].

A natural step in this area is an evaluation methodology that requires a deeper level of inference and of analysis of text.

TASK DESCRIPTION

The purpose of the task is to test systems' ability to understand the meaning communicated by a text. The focus is on semantic comprehension.

In contrast to text mining (or text harvesting, sometimes also called macro-reading), where the system reads and combines evidence from

[1] Lucy Vanderwende, Answering and Questioning for Machine Reading, American Association for Artificial Intelligence, March 2007

[2] Question Answering (QA): systems are designed to find exact answers to questions in a large collection of documents. A typical QA architecture is based on ranking, where the important thing is to answer the questions even though a system is not confident about the correctness of its answers.

[3] Recognizing Textual Entailment: system must decide whether the meaning of a text (called Text) entails the meaning of another text (called Hypothesis) - whether the meaning of the hypothesis can be inferred from the meaning of the text.

[4] Answer Validation Exercise (AVE): is defined as the task of deciding whether, given a question and an answer from a QA system, the answer is correct or not. The conclusions obtained from the results at AVE show that the addition of a validation module including more reasoning could contribute to the improvement of results in QA.

hundreds or thousands of texts, Machine Reading (MR) is the task of obtaining an in-depth understanding of just one – or a small number – of texts.

The setting is similar to the scenario of humans learning a new language and dealing with Reading Comprehension tests. As in such tests, the task focuses on the reading of single documents and identifying the answers to a set of questions. The identification of correct answers requires various kinds of inference and the consideration of previously acquired background knowledge.

In order to allow systems to acquire the same background knowledge needed for answering the tests, ad-hoc collections are being created – one for each of the topics - in all the various languages involved in the exercise.

Texts will be drawn from a diverse range of sources e.g.: newspapers, newswire, web, blogs, Wikipedia entries. Collections already used in previous CLEF campaigns will also be adopted.

The systems are required to answer each of the questions using the information contained in the Test Documents, with the help of additional knowledge obtained through the Background Collection provided. Although this background knowledge may be used to assist with answering the questions, the principal answer will be among the facts contained in the Test Document given.

The task is mainly focused on developing the reasoning capabilities of a system rather than on the acquisition of complex pieces of knowledge.

DATA SET

The data set is made up of a series of tests. Each test consists of one single document (Test Document) with several questions and a set of choices per question. So, the task is a Reading Comprehension test of the given document.

Systems will be given at least:

- 6 Test Documents (2 documents for each of the three topics)
- 10 questions per document with 5 choices for each question

These numbers represent the minimum amount that will be proposed – possibly more documents and questions may be present depending on feasibility.

Topics, documents and questions will be made available in *English, German, Italian, Romanian, and Spanish*. These materials will be exactly the same in all languages, created using parallel translations: all questions will be produced in English only and then translated into all the other languages of the task.

The Background Collections (one for each topic) are comparable (but not identical) topic-related collections created in all the different languages. The Background Collections will be available to all participants signing a license agreement. Thus, the learning of knowledge could be in one language or several.

QUESTIONS

- are in the form of multiple choice, where for each question a list of possible answers is given;
- focus on testing the comprehension of *one single document* for each reading test;
- test the *reasoning capabilities* of systems, which means that inferences, relative clauses, elliptic expressions, meronymy, metonymy, temporal and spatial reasoning, and reasoning on quantities may be exploited;
- involve *background knowledge*, i.e., information that is not present in the Test Document given. Questions will be posed in such a way that they cannot be answered without background knowledge.

Questions may be of the following types:

1. FACTOID: Where or When or By-Whom
2. CAUSAL: What was the cause/result of Event X?
3. HYPOTHETICAL: What would have happened if [not] X?
4. COMPOSITE: the question includes a second question whose answer is used to determine the final answer, for example, "How large is the city in which Princess Diana died?"

ANSWERS

- For each question, systems must identify the correct answer by selecting it from the five alternatives proposed:
  - there will always be one and only one correct option;
- Determining the answer will require a deep understanding of the text and not merely a mechanical repetition of the input:
  - the options are not substrings of the test document given;
  - all kinds of textual inferences may be needed, i.e., lexical (acronymy, synonymy, hyperonymy), syntactic (nominalization / verbalization, causative, paraphrase, active/passive), discourse (coreference, anaphora ellipsis);
- The direct and immediate answer will always be present in the text given for reading. But to recognize that it is the direct answer,

systems may need some background knowledge, which is contained in the Background Collection;

- Systems may be required to synthesize information from multiple sentences, as the material for answering the questions may be split throughout the Test Document and even the Background Collection;
  Systems are required to select the answer and also to provide the document(s) and sentence(s) that helped them (directly or indirectly) to identify the correct answer. Such provenance will not be used for formal evaluation, but for informal analysis and discussion;
- A system is not required to answer every question, as the C@1 measure is being used for evaluation (see below). Therefore, there are three possibilities:
  (1) To submit an answer and ask for it to be evaluated,
  (2) Not to submit an answer,
  (3) To submit an answer and ask for it not to be evaluated.
  Options (2) and (3) will result in identical C@1 scores but the differences could be interesting for further analysis (i.e. to provide additional feedback about systems self-validation). The Output Format section below gives more details.

RUNS

Participants are allowed to submit a maximum of 10 runs. The first run must be produced using nothing more than the knowledge provided in the Background Collections. Additional runs can include other sources of information, e.g. ontologies, rule bases, web, Wikipedia, etc., or other types of inferences. All resources used to acquire the knowledge must be listed in the submission file (see section Output Format).

FORMATS

The DTD for the input and output format can be downloaded from the QA4MRE website http://celct.fbk.eu/QA4MRE/

INPUT FORMAT

The test set will be formatted as an xml file (UTF-8 encoded). The xml will be structured with elements containing the following information:

topic t_id t_name doc_reading_pair dr_id doc d_id
reading-test r_id question q_id q_str answer a_id

where:

- **topic** includes the following elements: topic, questions, and multiple choice answers;
- **t_id** is the id of the topic (from 1 to 3);
- **t_name** is the title of the topic;
- **reading-test** includes the test document and all questions with the multiple choice answers which refer to the test document;
- **r_id** is the unique id of the reading test;
- **doc** is the test document which the questions are being asked again*st;*
- **d_id** is the id of the test documents;
- **question** is the question and the multiple choice answer;
- **q_id** is the id of the question (from 1 to 10);
- **q_str** is the question (UTF-8 encoded) string;
- **answer** is each of the five options of the multiple choice;
- **a_id** is the id of the answer (from 1 to 5).

i.e.:

```
<test-set>

<topic t_id="1" t_name="title of the topic">
          <reading-test r_id="1">
               <doc d_id="1">
                    Some Text here.......Some Text here....Some
                    Text here....Some Text here....Some Text
                    here....Some Text here.....Some Text
                    here.....Some Text here....Some Text here
               </doc>
               <question q_id="1">
                    <q_str> Text of question </q_str>
                    <answer a_id ="1">Text of answer </answer>
                    <answer a_id ="2">Text of answer </answer>
                    <answer a_id ="3">Text of answer </answer>
                    <answer a_id ="4">Text of answer </answer>
                    <answer a_id ="5">Text of answer </answer>
               </question>
               <question q_id="2">
                    <q_str> Text of question </q_str>
                    <answer a_id ="1">Text of answer </answer>
                    <answer a_id ="2">Text of answer </answer>
                    <answer a_id ="3">Text of answer </answer>
                    <answer a_id ="4">Text of answer </answer>
                    <answer a_id ="5">Text of answer </answer>
               </question>
               ...
               ...

               </question>
          </reading-test>
     .
     .
     .
     .
```

```
</topic>
</test-set>
```

OUTPUT FORMAT

A submission file must be an xml file (UTF-8 encoded) in the form:

```
<output run_id="XXXX1101XXXX">

<topic t_id="1" >
     <reading-test r_id="1">
          <question q_id="1" answered="YES">
               <answer a_id="1">
          <provenance doc_name="doc213">sentence text</provenance>
          <provenance doc_name="doc89">sentence text</provenance>
          <provenance doc_name="doc67"> sentence text</provenance>
          <provenance doc_name="doc67">sentence </provenance>
          ...
               </answer>
          </question>

          <question q_id ="2" answered="NO" />

          <question q_id ="3" answered="YES" >
               <answer a_id ="5">
                    <provenance doc_name="">sentence </provenance>
               </answer>
          </question>

          <question q_id ="4" answered="NO">
               <answer a_id ="1">
           <provenance doc_name="doc35">sentence</provenance>
               </answer>
          </question>

          <question q_id ="5" answered="NO"/>
               .
               .
               .
               .
               .
     </reading-test>
     .
     .
     .
     .
     .
     </topic>
</output>
```

where:

- **run_id** attribute of root element is an alphanumeric string which identifies the runs of each participant. It should be the concatenation of the following elements:

- ~ the **team ID** (sequence of four lower case ASCII characters),
- ~ the **current year** (11 stands for 2011),
- ~ the **number of the run** (01 for the first one, or 02 for the second one, and so on. The maximum number of runs which can be submitted by each system is 10),
- ~ the **language pairs** including both source and target languages,

Clearly, the content of this field never changes within the same submission file. Each submission file must be named, with an .xml extension, e.g. “clct11001itit.xml”;

- **topic** includes the following elements: topic, questions, and multiple choice answers;
- **t_id** is the id of the topic as given in the test-set (from 1 to 3);
- **reading-test** includes all the 10 questions and the multiple choice answers which refer to the test document;
- **r_id** is the id of the reading test as given in the test-set;
- **question** is the question and the multiple choice answers;
- **q_id** is the question number as given in the test set. Questions must be returned in the same ascending (increasing) order which appear in the test-set;
- **answered** indicates if the question has been answered or not (YES or NO must be set);
- **answer** is the correct option which answers the question. For each question, only one option (answer tag) must be selected.
  In the case in which the attribute “answered” is set to “NO” the system has two options:
  - o the “answer” element is not returned, meaning that the system prefers not to answer;
  - o return the candidate answer which system would have given;
- **a_id** is the id of the answer as given in the test-set;
- **provenance** includes the document(s) and sentence(s) relevant to identify the correct answer. For the second runs on, also additional resources used to acquire the knowledge must be listed here.
  doc_name is the name of the document where additional information has been found.

The provenance of information used, must be included in the following form: [doc-id in the doc_name attribute, relevant sentence fragment(s) as value of the provenance tag].
If the document is not in the background collection, systems should provide their own unique ID as follows: [their-corpus name in the doc_name attribute, relevant sentence fragment(s) as value of the provenance tag].
If no background knowledge is needed to answer the question, only the sentence(s) of the Text Documents must be specified, leaving

the doc_name slot empty.
If more than one document and sentence needs to be returned, the provenance element must be repeated as many times as the number of documents/sentences to be returned.
A maximum of 500 characters may be submitted as content of the provenance element.

EVALUATION

The evaluation of the output given by participant systems will be performed automatically by comparing the answers of systems against the gold standard collection with annotations made by humans. No manual assessment will be required.

The official measure of the QA4MRE evaluation campaign will be the C@1.[5] This measure, already tried in previous CLEF QA Tracks, rewards systems that, while maintaining the number of correct answers are able to reduce the incorrect ones by leaving some questions unanswered. The idea is to promote the development of a technology beyond the simple ranking of choices.

Each single test (set of questions related to a test document) will receive an evaluation score between 0 and 1 using c@1. Then, the average c@1 over all tests will be computed.

IMPORTANT DATES

Samples of tests	mid of February
Release of topics and background collections	March 31
Test set release	June 6
Run submissions	June 17
Results to the participants	July 1
Submission of Notebook Papers	August 15
CLEF Workshop	September 19-22, Amsterdam, The Netherlands

[5] Anselmo Peñas, Pamela Forner, Richard Sutcliffe, Alvaro Rodrigo, Corina Forascu, Iñaki Alegria, Danilo Giampiccolo, Nicolas Moreau and Petya Osenova. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In C. Peters, G. Di Nunzio, M. Kurimo, Th Mandl, D. Mostefa, A. Peñas, G. Roda. (Eds) *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, Revised Selected Papers. Lecture Notes in Computer Science 6241. Springer-Verlag, 2010.

Test questions will be posted on the QA4MRE website http://celct.fbk.eu/QA4MRE/ on June 6 and submissions will be due within 5 days from the first test set download and not later than June 17 by 11:59 p.m. (CEST). Late submissions will not be considered.

Participant runs will be submitted using an automatic submission procedure. Details about the submission procedure will be provided when the test data is released. Before completing the submission, a checking routine will be automatically run in order to detect format inconsistencies and common errors in the files (invalid document numbers, wrong formats, missing data, etc..). The automatic submission procedure will reject any run which is not compliant with the required format.

GENERAL COORDINATION

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

TECHNICAL COORDINATION

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

Technical Staff

- Corina Forascu, Alexandru Ioan Cuza University, Romania
- 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