Track Guidelines

INTRODUCTION / MOTIVATION

The task focuses on measuring the ability of systems to automatically understand texts. Reading comprehension tests are designed to measure how well human readers understand what they read. The methodology proposed in this evaluation campaign resembles reading comprehension and represents an evolution of previous evaluation approaches in Natural Language Processing (NLP), including Question Answering[1], Recognizing Textual Entailment[2], and Answer Validation[3]. However, the machine reading task requires a deeper level of text analysis, including inference. In order to comprehend a single text a large amount of background knowledge is needed to the point that the acquisition of this background knowledge is considered in itself a reading task. Thus, together with the tests, we provide large document collections that serve as a background for each particular reading test.

TASK DESCRIPTION

In contrast to text mining (or text harvesting, sometimes also called macro-reading), where the system reads and combines evidence from hundreds or even thousands of texts, Machine Reading (MR) is the task of obtaining an in-depth understanding of just one, or a small number, of texts.

As in the previous campaign, 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 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 organizers. Although the additional knowledge obtained from the background collection may be used to assist in answering the questions, the principal answer must be found among the facts contained in the source test documents provided.

---

[1] 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.

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

[3] Answer Validation (AV): 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 at the Answer Validation Exercise (AVE) show that the addition of a validation module including more reasoning could contribute to the improvement of results in QA.

BACKGROUND COLLECTIONS

In addition to the source text, systems are provided with a collection of additional texts on the same topic, from which they may acquire the reading capabilities and draw the knowledge, if needed, to overcome any knowledge gaps in the source text. Similarly to last year, ad-hoc collections will be provided in all the languages involved in the exercise. Capitalizing on the experience of last year, these collections will be created by crawling the web: the collections produced last year will be used for the 2012 campaign, with the addition of 1,000 new documents for each topic in all languages. These collections are thus comparable across languages but are not parallel. Texts will be drawn from many sources: newspapers, newswire, web pages, blogs and Wikipedia entries. Thus the kind of knowledge provided is generic with respect to each topic, containing for example the most common classes and instances, frequent assertions, and general relations between these assertions such as causality, etc.
The background collections will be made available to all participants at the beginning of April, subject to signing a licence agreement, so that they can be used to acquire domain specific knowledge —in one language or several— prior to taking part in the QA4MRE task.

The use of the background collections is highly recommended though not mandatory.

TEST DATA

The 2012 test set will be composed of 4 topics, namely "Aids", "Climate change" and "Music and Society" — the same topics adopted last year— plus one additional new topic, namely "Alzheimer". Although the pilot task will be concerned with the analysis of scientific or specialized language, the main task being described here will deal with generic or popular language. Each topic will include 4 reading tests. Each reading test will consist of one document, accompanied by 10 questions, each with a set of five answer options per question. So, for each language task, there will be in total:

- 16 test documents (4 documents for each of the four topics)
- 160 questions (10 questions for each document) with
- 800 choices/options (5 for each question)

LANGUAGES

Test documents, questions, and options will be made available in *English, German, Italian, Romanian, Spanish* and —new this year— *also Arabic* and *Bulgarian*. These materials will be exactly the same in all languages, created using parallel translations.

QUESTIONS

- are in the form of multiple choice, where for each question, 5 possible answers are given;
- focus on testing the comprehension of *one single document*;
- 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;
- may involve *background knowledge*, i.e., information that is not present in the test document given. In such cases, information from the Background collections is needed to fill in the knowledge gap to answer the question.

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. METHOD: How did X do Y? Or: In what way did X come about?
4. PURPOSE: Why was X brought about? Or: What was the reason for doing X?
5. WHICH IS TRUE: Here one must select the correct alternative from a number of statements, e.g. What can a 14-year-old girl do?

ANSWERS

- The system is not required to answer every question, as the C@1 measure will be 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 (e.g., to provide additional feedback about a system's self-validation). The Output Format section below gives more details.
- Systems may be required to synthesize information from multiple sentences, as the material for answering the questions may be distributed throughout the test document;
- The direct and immediate answer is always present in the test document. However, to recognize that it is the answer, systems may need some background knowledge; Various kinds of textual inferences may be needed, e.g., lexical (acronymy, synonymy, hyperonymy), syntactic (nominalization / verbalization, causative, paraphrase, active/passive), discourse (coreference, anaphora ellipsis), etc.

RUNS

Participants are allowed to submit a maximum of 10 runs. Each run must be categorized as one the following types, depending on the resources that have been used to assist in answering the questions:

- No external resource is used (only the test document);
- Only the test document and the associated background collection are used;
- The test document and other resources are used, but not the background collection;
- The test document together with both the background collection and other resources are used.

Participants can experiment with all the above possibilities, specifying the resources used in the submission file (see 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_name t_id reading-test r_id doc d_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 (a number 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 against;
- **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 (a number from 1 to 10);
- **q_str** is the question string (UTF-8 encoded);
- **answer** contains one of the five multiple choice options;
- **a_id** is the id of the answer (a number 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). The xml will be structured with elements containing the following information:

run_id topic t_id reading-test r_id q_id answered answer a_id other-resources resource

- The **run_id** attribute of the 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** (12 stands for 2012),
    - **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 number corresponding to the type of resources** used in this run to assist with answering the questions. i.e.:
        (1) No external resource (only the test document)
        (2) Only the test document with its associated background collection
        (3) The test document with other resources but not the background collection
        (4) The test document with both the background collection and other resources

      In cases 3 and 4 the "other-resources" tag must also be filled, specifying the list of resources used.
    - **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 given an .xml extension, e.g. "clct12014itit.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 10 questions relating to the test document, together with their multiple choice answers ;
- **r_id** is the id of the reading test as given in the test-set;
- **question** is the question and its 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 in which they appear in the test-set;
- **answered** indicates if the question has been answered or not (YES or NO must be set);
- **answer** contains the candidate answer, as determined by the system. For each question, only one option (answer tag) must be selected.
  In the case where the attribute "answered" is set to "NO" the system has two options:
    - the "answer" element is not returned, meaning that the system prefers not to answer;
    - The system returns in the "answer" element the candidate answer which the system would have given;
- **a_id** is the id of the answer as given in the test-set;
- **other-resources** specifies the list of resources used to answer the questions. This tag must be included only in the cases in which the type of resources used in the run is number 3 or 4.
- **resource** specifies the name of the resource used.

i.e.

```
<output run_id="XXXX1101YXXXX">
<topic t_id="1" >
     <reading-test r_id="1">
          <question q_id="1" answered="YES">
```

```
            <answer a_id="1"/>
        </question>

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

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

        <question q_id ="4" answered="NO">
            <answer a_id ="1"/>
        </question>

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

<other-resources>
    <resource>Name of resource</resource>
    <resource>Name of resource</resource>
    ...
    ...
</other-resources>
</output>
```

EVALUATION

Scoring of the output produced 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 performed.

Each test will receive an evaluation score between 0 and 1 using c@1[4]. This measure, already tried in previous CLEF QA Tracks, encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered.

Systems will receive evaluation scores from two different perspectives:

1. at the question-answering level: correct answers are counted individually without grouping them;
2. at the reading-test level: figures both for each reading test as a whole, and for each separate topic are given.

[4] Anselmo Peñas and Alvaro Rodrigo. A Simple Measure to Assess Non-response. In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA, June 19-24, 2011

IMPORTANT DATES

Release of background collections	April 6
Test set release	June 5
Run submissions	June 15
Individual results to participants	June 25
Submission of Working Notes Papers	August 17
CLEF Workshop	September 17-20, Rome, Italy

Test questions will be posted on the QA4MRE website http://celct.fbk.eu/QA4MRE/ on **June 5** and submissions will be due **within 5 days** from the first test set download and not later than **June 15** 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 automatically be 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 that does not comply 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

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