Evaluating Machine Reading Systems through Comprehension Tests

Anselmo Peñas[1], Eduard Hovy[2], Pamela Forner[3], Álvaro Rodrigo[1], Richard Sutcliffe[4], Corina Forascu[5], Caroline Sporleder[6]

1 NLP&IR group, UNED, Spain
2 ISI / USC, USA
3 CELCT, Italy
4 University of Limerick, Ireland
5 Al. I. Cuza University of Iasi, Romania
6 Saarland University, Germany

anselmo@lsi.uned.es, hovy@isi.edu, forner@celct.it, alvarory@lsi.uned.es, richard.sutcliffe@ul.ie, corinfor@info.uaic.ro, csporled@coli.uni-sb.de

Abstract

This paper describes a methodology for testing and evaluating the performance of Machine Reading systems through Question Answering and Reading Comprehension Tests. The methodology is being used in QA4MRE (QA for Machine Reading Evaluation), one of the labs of CLEF. We report here the conclusions and lessons learned after the first campaign in 2011.

Keywords: Question Answering, Machine Reading, Evaluation

1. Introduction

In contrast to traditional QA, where answers are determined by skimming large document collections, Machine Reading (MR) systems read only a handful of texts and analyse them in depth in order to obtain answers (Etzioni et al., 2006). Successful systems must generally perform complex inferences. Since frequently the source text is incomplete, the systems need access to background collections of documents or other sources of information such as Wikipedia or databases of facts.

Evaluating MR is complex. Since each system has its own internal knowledge representation, cross-system comparisons (before and after reading) are difficult. A simpler approach is to treat MR evaluation as an Information Extraction task (did the system extract the correct information?) or as a QA task (is the system able to read and reason in order to arrive at the correct answer?). In the QA approach, two models are possible. In one, a formal language (target ontology) is defined, and systems are required to translate texts into this representation. Evaluation can then take place by submitting structured queries in the formal language in order to determine if certain inferences have been made. However, there are certain problems with this approach and in consequence we have developed a new model, influenced by previous research on QA and reading comprehension (Wellner et al., 2006). A series of questions is asked for each document, and each question has a set of multiple-choice answers. This allows complex questions to be asked but makes evaluation simple and completely automatic. The evaluation architecture is completely multilingual: test documents, questions, and their answers are identical in all the supported languages. Background text collections are comparable collections harvested from the web for a set of predefined topics. This approach allows deep natural language processing issues to be investigated in both monolingual and cross-lingual contexts.

2. Motivation

For some years it has been clear that there is an upper bound of 60% accuracy in QA systems, despite more than 80% of questions being answered correctly by at least one participant. Analysis uncovered a problem of error propagation in the traditional QA pipeline (Question Analysis, Retrieval, Answer Extraction, Answer Selection/Validation). Thus, in 2006 we proposed a pilot task called the Answer Validation Exercise (AVE) (Peñas et al., 2006). The aim was to produce a change in QA architectures, giving more responsibility to the validation step, which could help to overcome the limitations of pipeline processing.

After three AVE campaigns, we transferred our conclusions to the main QA task at CLEF in 2009 and 2010 (Peñas et al., 2010). The first step was to introduce the option of leaving questions unanswered, which is a strategy related to the development of validation technologies. We needed a measure able to reward systems that withheld answers to certain questions if they were not sure of them. The result was c@1 (Peñas and Rodrigo, 2011), a measure tested during the 2009 and 2010 QA campaigns at CLEF, and also used in the current evaluation.

However, this change was not enough. Almost all systems continued using IR engines to retrieve relevant passages and then tried to extract the answer from that. This was not the change in architecture we expected, and again, results remained below the 60% pipeline upper bound. We concluded that the change in architecture requires a

previous development of answer validation/selection technologies. For this reason, in the current formulation of the task, the step of retrieval is put aside for a while, in place of a focus on the development of technologies able to work with a single document.
This development parallels the introduction in 2009 of the Machine Reading Program (MRP) by DARPA in North America. The goals of the program are to develop systems that perform deep reading of small numbers of texts in given domains and to answer questions about them. Analogously to QA4MRE, the MRP program involves batteries of questions for the evaluation of system understanding. However, testing queries were structured according to target ontologies, forcing participant teams to focus on the problem of document transformation into the formal representation defined by these target ontologies. Thus the Machine Reading challenge had to pass through the Information Extraction paradigm.
In QA4MRE we think it is important to leave the door open to find synergies with emerging research areas such as those related to Distributional Semantics, Knowledge Acquisition, and Ontology Induction. For this reason, we are agnostic with respect to the query language and the internal machine representation. Thus, questions and answers are posed in natural language.

3. The QA4MRE Task

As we have seen, the task this year was to answer a series of multiple choice tests, each based on a single document (Peñas, Hovy et al., 2011). Tests comprised three topics, namely "Aids", "Climate change", and "Music and Society". Each topic included 4 reading tests. Each reading test consisted of a single document, with 10 questions and a set of five choices per question. In this campaign, the evaluation had in total:

- 12 test documents (4 documents for each of the three topics),
- 120 questions (10 questions for each document) with
- 600 choices/options (5 for each question).

3.1 Test Documents

After some consideration, we used parallel documents, in English, German, Italian, Romanian, and Spanish, taken from the Technology, Entertainment, Design (TED) conferences (www.ted.com). Each TED event consists of a series of twenty-minute presentations by prestigious speakers, from fields such as politics, entertainment and industry. The selected talks range in length between 1,125 and 3,580 words. We verified that the translations (based on English transcriptions) are of very high quality.

3.2 Questions

Questions were posed by studying the test documents, as is the norm in QA evaluations. Questions may refer to:

- facts that are explicitly present within a single sentence in the text,
- facts that are explicitly present, spread over several sentences,
- facts that are not explicitly mentioned, but are one inferential step away (cf. the RTE challenge),
- facts that are explicitly mentioned but require some inference in order to be connected together so as to form the answer.

Out of the 120 questions, 44 needed extra information from the background collection, while the document alone was sufficient for 76. 38 questions had the answer in the same paragraph, while for 82, several paragraphs were needed. Questions were posed so that answers were not merely a mechanical repetition of the input. Instead, all kinds of textual inferences could be requested, such as *lexical* (acronym, synonymy, hyperonymy-hyponymy), *syntactic* (nominalization-verbalization, causative, paraphrase, active-passive), and *discourse* (co-reference, anaphora ellipsis).

3.3 The Background Collections

One focus of the task is the ability to extract different types of knowledge and to combine them as a way to answer the questions. In order to allow systems to acquire the same background knowledge, ad-hoc collections were created — one for each of the topics — in all the languages involved in the exercise, i.e., English, German, Italian, Romanian, and Spanish. These collections were created by crawling the web. They are thus comparable across languages but are not parallel. The collections were made available to all participants at the beginning of April so that they could be used to acquire domain specific knowledge — in one language or several — prior to taking part in the QA4MRE task.

3.4 Evaluation

Evaluation was performed automatically by comparing the answers given by systems to the ones prepared by the organisers. No manual assessment was required because of the multiple-choice format.
Each test received an evaluation score between 0 and 1 using c@1. This measure encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered.
The task allowed us to evaluate systems from two different perspectives:
- As a question-answering evaluation, where we just counted correct answers without grouping them.
- As a reading-test evaluation, where we obtained figures both for each reading test, and for each topic.
Concerning a baseline level of performance, there are five possible answers to each question, with exactly one being correct. Assuming that all questions are answered, the random baseline is 0.2 (both for accuracy and c@1).

3.5 Participation and Results

Out of the 25 groups that originally registered, 12 participated in the task, submitting 62 runs in 3 different languages (German, English, and Romanian). All runs were monolingual; no team attempted a cross-language task.
Participants were allowed to submit a maximum of 10 runs. The first run was to be produced using nothing more than the knowledge provided in the background collections. Additional runs could include other sources of information, e.g., ontologies, rule bases, the web, Wikipedia, etc., or other types of inferences.

As for system performance at the question-answering evaluation level, only one team (jucs) scored above 50%. From a reading test perspective, no group passed. System performance was not significantly better than the random baseline. There is thus great potential for future improvement on both sides, not only system development but also in the evaluation methodology.

4 Lessons Learned

The first difficulty we had to address was the definition of Background Knowledge. This is required in order to make consistent decisions about the methodology for harvesting background collections and for developing the testing questions and answers

4.1 Definition of Background Knowledge

Reading Comprehension tests are routinely used to assess the degree to which people comprehend what they read, so we work with the hypothesis that it is reasonable to use these tests to assess the degree to which a machine "comprehends" what it is reading.
When reading a text, a human performs two processes, namely:

1. s/he partially/fully understands its meaning;
2. if needed, s/he makes additional inferences from the text, i.e., performs some kind of reasoning, and solves the textual inferences (linguistic/lexical, co-reference), using previously acquired experience/knowledge of any type.

We assume that the answer to a question almost always requires some prior knowledge. Resources such as wordnets, framenets, paraphrase bases, knowledge bases are aimed at make different kinds of prior knowledge available for the machine.
We add to these resources the possibility to acquire background knowledge from a large collection of related documents. The advantage is the opportunity to gather probability distributions linked to knowledge, and explore distributional approaches.
The answers to the questions should never come from this prior knowledge alone. The answer must be found in the Test Documents, but references to information outside it may be required, as there may be explicit and implicit references to entities, events, dates, places, situations, etc. pertaining to the topic.
Therefore, the definition of Background Knowledge must be given, in our case, in terms of the relation between the testing questions and answers, and the background collection. In other words, what is the use of the prior knowledge? For this purpose, we distinguish at least four main types of background knowledge (although in fact it is a continuum):

1. Very specific facts related to the document being read. For example, the relevant relation between two concrete people involved in a specific event.
2. General facts not specific to any particular event. For example, geographical knowledge, main players in international affairs, movie stars, world wars. Also acronyms, transformations between quantities and measures, etc.
3. General abstractions that humans use to interpret language, to generate hypotheses, or to fill missing or implicit information. For example, abstractions as the result of observing the same event with different players (e.g., petroleum companies drill wells, quarterbacks throw passes, etc.)
4. Linguistic knowledge. For example, synonyms, hypernyms, transformations such as active/passive, or nominalizations. Also transformations from words to numbers, meronymy, metonymy.

Obviously this is not an exhaustive list. For example, we are not talking about ontological relations that enable temporal and spatial reasoning, or reasoning on quantities.
In summary, the questions should be answerable by most humans using their general knowledge, without the need to explore a specific document of the background collection. Examples of inferences we allow are:

1. Linguistic inferences such as coreference, deictic references (like "then" and "here"), etc.)
2. Simple ontological inferences such as considering part-of relations or obtaining direct super-concepts for common objects.
3. Inferences considering causal relations or procedural steps in "life scripts" like visiting a restaurant or attending a concert.
4. Inferences that require composing several answers, in particular answering one part of the question using the background collection and then, with its answer, answering the other part of the initial question (e.g., "*Who is the wife of the person who won the Nobel Peace Prize in 1992*").

4.2 Creation of Background Collections

This is a very important element of the evaluation setting. It connects the task with research in Information Retrieval. The goal of reference/background collections is to contextualize the reading of a single document related to the topic. Thus, we could expect that in the future this step could be done on the fly as a retrieval process given the single text being read.
But for now, the organization is doing this for two main reasons: to enable better comparison among participant systems, and to focus on the Reading Comprehension problem. Therefore, it is very important to develop a good methodology to build these background collections for the evaluation task.
Ideally, the background collection should completely cover the corresponding topic. This is sometimes feasible and sometimes not. For example, in the case of the pilot task on Biomedical documents about Alzheimer's disease (see Section 5.3), a set of experts built a query (a set of conjunctions and disjunctions over 18 terms) that closely approximates the retrieval of all relevant documents (more than 66,000) without introducing much noise.
However, this is not so easy in more open domains such as Climate Change, or when one wants to consider non-specialized sources of information. In these cases, we crawl the web using, for each language and topic, a list of keywords and a list of sources.
Keywords are translated into English and then translated into the other languages. Documents may be crawled from a variety of sources: newspapers, blogs, Wikipedia,

journals, magazines, etc. The web sources are obviously language dependent, so each language requires also a list of possible web sites with documents related to the topic.
However, we face the same problem as does traditional Information Retrieval: we want all relevant documents (and only them), and we use queries (keywords) to retrieve them.
We realized that, since the organizers knew the test set, they used that information to select the keywords, and ensure the coverage of the questions. The effect is not only that background collections don't cover completely the topic, but also that the collection has some bias with respect to the real distribution of concepts.
The assumption that the ideal background collection should include all relevant documents for the topic is explicit, and we organizers have it in mind. Our first strategy with the aim of ensuring the coverage of the topic as much as possible is to make the topic specific enough (e.g., *AIDS medicaments* instead of *AIDS*). The second strategy is to try to cover (at least partially) each of the possible *"dimensions/aspects"* of that topic, which we do as follows: First, we locate a good central overview text, such as a Wikipedia article that "defines" the topic, "suggests" its principal aspects (often in subsections), and provides some links. Then, we enumerate these dimensions and prepare a set of queries for each dimension. We document this process with three benefits: (i) to know what organizers and participants can expect or not from the collection; (ii) to give another dimension of re-usability; and (iii) to explore how Machine Reading will connect to Information Retrieval in the future.

5 Toward 2012

After the 2011 QA4MRE evaluation, we have prepared for the next campaign by refining the methodological issues above. In 2012, we will have a main task and two pilot exercises.

5.1 Main Task

The main task will remain the same for participants. Background collections, test documents, and reading tests will be available in Arabic, Bulgarian, English, German, Italian, Romanian, and Spanish. In addition to last year's topics (AIDS, Climate Change, Music and Society), we will include a topic on Alzheimer's disease. This new topic is related to a new pilot on Biomedical texts. The difference is that the reference collection for the main task is built from general public sources and for the pilot the source is the PubMed repository.
Having these two parallel exercises on the same topic but in different domains opens the door to evaluating research on the challenges of domain and language adaptation, the use of knowledge in one domain captured in the other, the differences in the background knowledge acquired, the differences between questions and answers in each domain, etc.

5.2 Pilot on Processing Modality and Negation for Machine Reading

This exercise is aimed at evaluating whether systems are able to understand extra-propositional aspects of meaning, such as 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.
For this purpose, participants will receive some tests where they have to decide whether some events are Asserted, Negated, or Speculative. Our plan is to integrate modality and negation in the main task next year.

5.3 Machine Reading on Biomedical Texts about Alzheimer's disease

This Pilot is aimed at setting questions in the Biomedical domain with a special focus on one disease, namely Alzheimer's. 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 66,222 Medline abstracts. 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.

6 Conclusions

In 2011, the QA@CLEF task was characterised by two major innovations. First, there was a transition from traditional Question Answering based on shallow analysis of large document collections to a new focus involving deep analysis of individual documents. Over the years, the QA challenges adopted simple questions that required almost no inferences to find the correct answers. These surface-level evaluations promoted QA architectures based on Information Retrieval (IR) techniques, in which the final answers were obtained after focusing on selected portions of retrieved documents and matching sentence fragments or sentence parse trees. No real understanding of documents was achieved, since none was required by the evaluation. Machine Reading, on the other hand, requires the automatic understanding of texts at a deeper level, so this task encourages participants to build a different kind of system.
The second innovation of the task lay in the evaluation. Instead of manually inspecting answers to judge whether they were correct, evaluation was entirely automatic. This was made possible by adopting questionnaires comprising multiple-choice questions whose exact answers could be determined in advance. This strategy also enabled more complex types of question to be asked as well as posing fewer restrictions on the form of the answers.
Significant lessons were learned from this new evaluation which was also well received by the QA community. This opens the way for future evaluations based on similar principles.

7 Acknowledgements

Anselmo Peñas and Álvaro Rodrigo have been partially supported by the Research Network MA2VICMR (S2009/TIC-1542) and Holopedia project (TIN2010-21128-C02).
Eduard Hovy was partially supported in DARPA's Machine Reading Program under contract number: FA8750-09-C-0172.
Caroline Sporleder is supported by the German Research

Foundation, DFG (Cluster of Excellence on “Multimodal Computing and Interaction”, MMCI).
Pamela Forner was partially supported by the PROMISE Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

8 References

Etzioni, O., Banko, M., and Cafarella, M.J. (2006). Machine Reading. In *Proceedings of the 21st National Conference on Artificial Intelligence*.

Peñas, A., and Rodrigo, Á. (2011). 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).*

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á, Sutcliffe, R., Forascu, C., and Sporleder, C. (2011). Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. *CLEF 2011 Labs and Workshop Notebook Papers*, Amsterdam, 19-22 September, 2011. ISBN 978-88-904810-1-7, ISSN 2038-4726.

Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, Á., Forascu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P. (2010). 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 Vol. I Text Retrieval Experiments*. CLEF 2009 Revised Selected Papers. LNCS 6241. Springer-Verlag.

Peñas, A., Rodrigo, Á., Sama. V., Verdejo, F. (2006). Overview of the Answer Validation Exercise. In C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval.* 7th Workshop of the Cross-Language Evaluation Forum. Revised Selected Papers.

Wellner, B., Ferro, L., Greiff, W., and Hirschman, L. (2006). Reading Comprehension Tests for Computer-based Understanding Evaluation. *Natural Language Engineering* 12(4), 305–334.