QA4MRE 2011-2013: Overview of Question Answering for Machine Reading Evaluation

Anselmo Peñas[1], Eduard Hovy[2], Pamela Forner[3], Álvaro Rodrigo[1], Richard Sutcliffe[4], and Roser Morante[5]

[1] NLP&IR Group, UNED, Spain
{anselmo,alvarory}@lsi.uned.es
[2] Carnegie Mellon University, USA
hovy@cmu.edu
[3] CELCT, Italy
forner@celct.it
[4] School of CSEE, University of Essex, UK
rsutcl@essex.ac.uk
[5] CLiPS, University of Antwerp, Belgium
roser.morante@ua.ac.be

Abstract. This paper describes the methodology for testing the performance of Machine Reading systems through Question Answering and Reading Comprehension Tests. This was the attempt of the QA4MRE challenge which was run as a Lab at CLEF 2011–2013. The traditional QA task was replaced by a new Machine Reading task, whose intention was to ask questions that required a deep knowledge of individual short texts and in which systems were required to choose one answer, by analysing the corresponding test document in conjunction with background text collections provided by the organization. Four different tasks have been organized during these years: Main Task, Processing Modality and Negation for Machine Reading, Machine Reading of Biomedical Texts about Alzheimer's disease, and Entrance Exams. This paper describes their motivation, their goals, their methodology for preparing the data sets, their background collections, their metrics used for the evaluation, and the lessons learned along these three years.

1 Introduction

The general goal of the Question Answering for Machine Reading Evaluation (QA4MRE) is to assess the ability of systems in two reading abilities: to answer questions about a text under reading, and to acquire knowledge from reading, especially the knowledge involved in the textual inferences that bridge the gap between texts, questions and answers.

The evaluation of these abilities can be approached in two principal different ways: the first one is to define a formal language (e.g., relational database), ask the systems to translate texts into the formal language representation (i.e., Information and

P. Forner et al. (Eds.): CLEF 2013, LNCS 8138, pp. 303–320, 2013.

Relation Extraction), and then evaluate systems by using structured queries formulated in the formal language.

The second main approach is agnostic with regard to any particular representation: systems' input queries about the text are natural language questions. This is related to how Question Answering (QA) is being articulated during the last decade. In QA4MRE we follow this approach but with a significant change with respect to previous QA campaigns over unstructured text.

1.1 From QA to Reading Comprehension Tests

By 2005 we realized that there was an upper bound of 60% of accuracy in system performance, despite more than 80% of the questions being answered by at least one participant. We understood that we had a problem of error propagation in the traditional QA pipeline (Question Analysis, Retrieval, Answer Extraction, Answer Selection/Validation). Thus, in 2006 we proposed a task called Answer Validation Exercise (AVE) [6]. The aim was to produce a change in QA architectures to give more responsibility to the validation step. In AVE we assumed there was a previous step of hypothesis generation and the hard work had to be done in the validation step. This is a kind of classification task that could take advantage of Machine Learning. The same idea is behind the architecture of IBM's Watson (DeepQA project) that successfully participated at Jeopardy [1].

After the three editions of AVE we tried to transfer our conclusions to the main QA task at CLEF 2009 and 2010 [9]. The first step was to introduce the option of leaving questions unanswered. This is an easy way of testing systems' confidence: if a system is not sure about its answers, it can decide to let unanswered a question instead of risking giving an incorrect answer. This is related to the development of validation technologies. Then, we needed a measure able to reward systems that reduce the number of questions answered incorrectly without affecting system accuracy, by leaving unanswered the questions they estimated they couldn't answer. The measure was an extension of accuracy called c@1 [5], tested during 2009 and 2010 QA campaigns at CLEF, and used also in the current evaluation.

However, this change wasn't enough. Almost all systems continued relying on IR engines to retrieve relevant passages and then trying to extract the exact answer from them. This is not the change in the architecture we expected, and again, results didn't go beyond the 60% pipeline upper bound. Finally, we understood that the change in the architecture requires to put more effort on the 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, focusing on the development of technologies able to work with a single document, and to answer questions about it.

In the new setting, we started again de-compounding the problem into hypothesis generation and validation. Thus, in QA4MRE we test systems only for the validation step. Together with the questions, the organization provides a set of candidate answers. This gives the evaluation the format of traditional Multiple Choice Reading Comprehension tests.

This development parallels the introduction in 2009 of the Machine Reading Program (MRP) by DARPA in North America. The goals of the program were 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 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 machine internal representation. Thus, questions and answers are posed in natural language.

1.2 Hypotheses, Research Questions and Specific Goals

Summing up, these are the hypotheses we make:

- Progress on Question Answering requires new architectures based on Hypothesis Generation and Answer Validation.
- There is a gap between texts, questions, and answers that requires, among other things, background knowledge and textual inference.
- Knowledge Bases of factual relations are not enough as sources of knowledge. Language interpretation requires other kinds of knowledge attached to language in different layers, from paraphrases to common sense general axioms.

Then, several research questions arise, including:

- What is the role of knowledge in bridging the gap between Texts, Questions, and Answers? To what extent can this knowledge be automatically derived from large text collections?
- What kind of synergies can be found between the use of relational knowledge bases, distributional semantics, and propositional semantics?
- Are systems able to consider extra-propositional aspects of meaning like modality and negation?
- How can one determine systems' levels of inference?
- What benchmarks best measure future progress in the field?
- How to evaluate systems ability to ensure that an answer is correct or even, to decide that there are no correct answers among candidates?

The evaluation campaigns aimed at giving, at least, partial answer to those questions by means of developing an evaluation methodology with 100% reusable benchmarks able to measure progress in the future (in several languages). Once this task is accomplished, the task now is to determine the current state of the art, and envisage next steps in the research agenda.

1.3 Roadmap

In 2011 we defined the following principles and roadmap:

1. Focus on validation: Questions have attached a set of candidate answers.
    a. Step 1. All questions have one and only one correct candidate answer.
    b. Step 2. Introduce questions that require inference.
    c. Step 3. Introduce questions with no correct candidate answer.
    d. Step 4. Introduce questions that require textual inference after reading a large set of documents related to the test.
2. Introduce hypothesis generation: Organization provides reference collections of documents related to the tests.
    a. Step 5. Questions about a single document, but no candidate answers are provided.
    b. Step 6. Full setting of QA where systems have to generate hypotheses considering the reference collection and provide the answer together with the set of documents that support the answer.

After three years, we have addressed most of the first phase (Steps 1–4), but the question now is if systems have achieved performance levels that ensure a qualitative difference if we try phase 2.

2 The Task

The QA4MRE task focuses on the reading of single documents and the identification of the answers to a set of questions. Questions are in the form of multiple choice, each having several options, and only one correct answer. The detection of correct answers might eventually require various kinds of inference and the consideration of previously acquired background knowledge from reference document collections. 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. Thus, reading comprehension tests do not require only *semantic understanding* but they assume a *reasoning process* which involves using implications and presuppositions, retrieving the stored information, performing inferences to make implicit information explicit. Many different forms of *knowledge* take part in this process: linguistic, procedural, world-and-common-sense knowledge. All these forms coalesce during processing and it is sometimes difficult to clearly distinguish and reconstruct them in a system that needs additional knowledge and inference rules in order to understand the text and to give sensible answers.

By giving only a single document per test, systems are required to understand every statement and to form connections across statements in case the answer is spread over more than one sentence. Systems are requested to (i) understand the test questions, (ii) analyse the relation among entities contained in questions and entities expressed by the candidate answers, (iii) understand the information contained in the documents, (iv) extract useful pieces of knowledge from the background collections, (v) and select the correct answer from the five alternatives proposed.

From 2011 until 2013, four tasks have been organized in QA4MRE. These tasks are described in detail in the CLEF Online Working Notes.

2.1 Main Task

The main task has been available in several languages (including Arabic, Bulgarian, English, German, Italian, Romanian, and Spanish). Test sets were divided into topics (AIDS, Climate Change, Music, Society and Alzheimer's disease). For each topic a background collection was provided, together with a set of testing documents for which questions were formulated, and candidate answers offered [7, 8, 10].

The resulting benchmark contains parallel tests into several languages (documents, questions and candidate answers are translations), and comparable documents as background reference collections.

Questions were made by task organizers to test a pre-selected set of question types and different levels of inference. In many cases, selecting the correct answer requires to gather previous information from the reference collection.

2.2 Machine Reading on Biomedical Texts about Alzheimer's disease

This pilot task explored the ability of a system to answer questions using scientific language. The test posed questions in the Biomedical domain with a special focus on one disease, namely Alzheimer's. Texts were taken from PubMed Central related to Alzheimer's and from 66,222 Medline abstracts [4, 12].

Here, the specific domain enabled us to explore Machine Reading linked to controlled vocabularies, entity types, and a predefined set of relations among these entity types. Thus, the task aimed at finding contact points with approaches based on Information Extraction.

2.3 Japanese University Entrance Exams

In all previous tasks, questions were posed by organizers with the aim of evaluating automatic systems under different reading abilities, types of questions, inference degree, etc. However, these questions were developed for the task and, thus, they can be arguably artificial.

In the challenge of "Entrance Exams", the goal is to test systems in a real scenario, like in a Turing test. Thus, systems were evaluated under the same conditions humans are evaluated to enter the University of Tokyo. For this purpose, some exercises about Reading Comprehension were extracted from actual exams [13].

This exercise was organized in coordination with the "Entrance Exams" task at NTCIR. Exams were created by the Japanese National Center for University Admissions Tests and the "Entrance Exams" corpus was provided by NII's Todai Robot Project and NTCIR.

2.4 Processing Modality and Negation for Machine Reading

This task was aimed at evaluating whether systems were able to understand extra-propositional aspects of meaning like modality and negation [2, 3]. Modality is a

grammatical category that expresses aspects related to the attitude of the speaker towards his/her statements, including certainty, factuality, and evidentially. Negation is a grammatical category that allows changing the truth value of a proposition. Modality and negation interact to express extra-propositional aspects of meaning. This task exploited the same topics and background collections of the Main Task. However, test documents were specifically selected to ensure the properties required for the questions. Participating systems had to decide whether given events in the texts were Asserted, Negated, or Speculated. The task was offered in English only in 2011 and 2012. In 2013 we integrated modality and negation into the Main Task by including some questions that required this kind of processing in order to answer correctly.

3 The Background Collections

Human language text does not include all the information we want to transmit. This is because we omit information we know the reader will obtain from the context and their own language of the world. However, this fact represents a big issue for systems aimed at managing the knowledge contained in tests

Therefore, the use of Background Knowledge represents a very important element of the evaluation setting. Since no text is ever complete, the goal of reference/background collections is to contextualize the reading of a single document within its general topic, allowing systems to construct models of knowledge and inference as needed to overcome gaps, omissions, assumptions, and otherwise incomplete information in the given texts and questions. Such models can be constructed before the actual test or at run-time, at the discretion of the system.

We define *background knowledge* in terms of the relation between the testing questions (and answers) and the background collection. To determine the potential kinds of uses of prior knowledge, we distinguish at least four main types of background knowledge (although in fact it's a continuum):

1. Very specific facts related to the document under study. 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 such 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, and metonymy.

Obviously this is not an exhaustive list. For example, we do not include ontological relations that enable temporal and spatial reasoning, or reasoning on quantities, which are also all relevant. Nonetheless, we believe that the collections allow systems to

extract, formalize, and apply during QA processing a lot of the kinds of information that people call 'commonsense and world knowledge'.

It is important to develop a good methodology for building background collections for the evaluation task. Ideally, the background collection should cover completely the corresponding topic. This is feasible sometimes and unrealistic at others. For example, in the case of the pilot on Biomedical documents about Alzheimer's disease, a set of experts built a query (a set of conjunctions and disjunctions over 18 terms) that approximates very much the retrieval of all relevant documents (more than 66,000) without introducing much noise. However, this is not so easy in more open domains (e.g., Climate Change) or cases with 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, and each language also requires a list of possible web sites with documents related to the topic.

We realized after the first campaign that, since we organizers knew the test set, we used that information to select the keywords, and ensure the coverage of the questions. The effect is not only that background collections didn't cover completely the topic, but also that the collections have some bias with respect to the real distribution of concepts.

For this reason, the assumption that the ideal background collection should include all relevant documents for the topic (and only them) was made explicit, and as organizers we bear it in mind. Thus, we face the same problem as traditional Information Retrieval: we want all relevant documents (and only them), and we use queries (keywords) to retrieve them

The 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 rather than AIDS). The second strategy is to try to cover (at least partially) each of the possible principal 'dimensions/aspects' of that topic. How? First, by detecting a good central overview text, such as a Wikipedia article that defines the topic, 'suggests' its principal aspects, and provides links to additional good material. Then, organizers enumerate these dimensions and prepare a set of queries for each dimension. They 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.

4 Test Collections

The methodology developed for creating test collections translated into several languages consists of the following steps:

1. Four English documents are selected for each of the four topics (Aids, Alzheimer's, Climate Change, Music and Society). They are selected from

various sources and comprised the test documents against which questions were asked. Documents are chosen from copyright-free sources or by kind permission to the owners (as for example in 2013 with documents of the Editor in Chief, Editor and Oxford University Press).

2. In order to have a set of identical questions for the languages involved, test documents are translated by expert translators recruited form the Translation for Progress[1] platform for all languages.

3. To ensure that translations are faithful to the original document in both meaning and style and of good quality, all the documents are manually checked and corrected when necessary. We wanted to avoid a situation where portions of the original English text were left out of the translation in a particular target language, or perhaps modified or interpreted in a particular manner which would have made the question impossible to answer in that language.

4. Fifteen multiple-choice questions are then devised for each test document (the 'Main' questions). A question always had five candidate answers from which to choose, with one clearly correct answer and four clearly incorrect answers. The last edition included in all cases the fifth candidate answer "None of the above", and six of the fifteen questions were composed so as to have no answer in the text. The correct response to each of these six questions was thus "None of the above".

5. In addition to the fifteen Main questions, the 2013 edition included also one or more Auxiliary questions. Each Auxiliary question was a simplified version of an existing Main question. The format of these questions was identical to that of Main questions, i.e. a question followed by five multiple-choice answers. In most cases, the Auxiliary question required less inference to answer. The idea was that if a system was able to answer the Auxiliary question but not the corresponding Main question, the problem could be its ability to perform the missing inference.

6. Once the questions had been composed in the language of the original author, each was then translated into English. The English versions of the questions and candidate answers are carefully checked by a referee to verify that they are clear, that the intended answer is clearly correct, that the intended answer is in the test document, and that the other candidate answers are clearly incorrect. Questions are modified accordingly.

[1] `http://www.translationsforprogress.org/main.php` A Translation Exchange site linking volunteer translators (e.g., linguistics students or professionals in foreign languages interested in building experience as translators can link up with low-budget organizations who are in need of translation work, but without the budget to pay for it. There are currently over 1450 registered volunteer translator members (for 13 language combinations) and over 160 organization members. Translation for Progress database is open for viewing for the general public, but if you wish to post your profile or contact a volunteer translator, a registration is required.

7. The English versions are then used to translate each question into each of the languages of the task. The same process is used to translate each candidate answer (five per query).

8. The result of this process is a set with 240 Main questions and, in 2013, 44 Auxiliary questions in different languages, each with five multiple-choice answers. The final step is to check that the answer to each question was in fact present in the test document for all the languages of the task.

4.1 Questions

Questions covered five different question types: purpose, method, causal, factoid, and which-is-true. Factoid questions were divided into the following sub-types: Location, Number, Person, List, Time and Unknown. Examples of the basic question types are given below. We took care to spread the question types evenly for a given test document, aiming for two questions per type. Example questions:

PURPOSE: What is the aim of protecting protein deposits in the brain?
METHOD: How can the impact of Arctic drillings be reduced?
CAUSAL: Name one reason why electronic dance music owes a debt to Kraftwerk.
FACTOID (number): What is the approximate number of TB patients?
WHICH-IS-TRUE: Which problem is similar in nature to global warming?

For all questions, the direct answer was contained in the test document; however answering the questions typically required some background knowledge and some form of inference. The required knowledge could be linguistic or could involve basic world knowledge. Linguistic knowledge concerns, for example, the ability to perform co-reference resolution or detect paraphrases on the lexical or syntactic level. World knowledge has to be inferred from the background collection. For instance, the text might mention *Barack Obama* while the question might refer to *the first African American President.* The fact that Barack Obama is the first African American President needs to be learnt from the background collection in order to be able to answer the question.

Typical types of world knowledge involve, for instance, knowledge about the basic referents in a text, e.g., being aware that *Yucca Mountain* is in Nevada. Another type of world knowledge involves knowledge of "life scripts" such as "visiting a restaurant". Finally, the inference required can also be complex, involving several steps. For example, answering a question might require combining knowledge from the background collection with knowledge from the test document itself. For instance, the question "Who is the wife of the person who won the Nobel Peace Prize in 1992?" contains two facts P and Q, where P="wife of Y=?" and Q="winner of Nobel Peace Prize in 1992=Y". The latter information can be gleaned from the background collection whereas the former is contained within the test document itself.

For each test document, we aimed for a combination of simple, medium, and difficult questions. At most six questions per document did not require knowledge from

the background collection. Two of these were simple questions, i.e., the answer and the fact questioned could be found in the same sentence in the test document. Four questions were of intermediate difficulty in that the answer and the fact questioned were not in the same sentence and could, in fact, be several sentences apart. Finally, the remaining four questions did require utilizing information from the background collection. While not all question types require inference based on the background collection, all of them required some form of textual and linguistic knowledge, such as the ability to detect paraphrases, as we made an effort to re-formulate questions in such a way that the answers could not be found by simple word overlap detection. For each question, we kept track of the inference required to answer it. This made it easier to ensure that that inference could in fact be drawn on the basis of the background collection, i.e., that the background collection did indeed contain the relevant fact. It also makes it possible to carry out further analyses regarding which questions or types of questions were difficult for the systems and why.

When creating the questions, we took care not to introduce any artificial patterns that would help finding the correct answer. Thus we ensured that all answer choices for a question were approximately the same length and consistent with respect to formulation and content, that all of the wrong answers were plausible, and that the placement of the correct answers was random and balanced.

5 Evaluation

One of goals of QA4MRE is to promote a change in QA architectures giving more importance to the validation step over the IR component in order to improve results. This is why we consider the possibility of leaving questions unanswered. The idea is that systems might reduce the amount of incorrect answers while keeping the proportion of correct ones, by leaving some questions unanswered.

Then, given a question with its corresponding candidate answers, a participant system can return two kinds of responses:

- An answer selected from the set of candidate ones for that question
- A *NoA* answer. This response should be given if the system considers it is not able to find enough evidences about the correctness of candidate answers and it prefers not to answer the question instead of giving an incorrect answer. Moreover, the system can return as a hypothetical answer the candidate one that it would have been selected, which allows to give some feedback about its validation performance.

The assessments of system's responses are given automatically by comparing them against the gold standard collection. Therefore, no manual assessment was required, which reduces the effort of the evaluation once the collections have been created and makes easier the future development of systems. Each system's response receives one and only one of the following three possible assessments:

- *Right* if the system has selected the correct answer among the set of candidate ones of the given question;
- *Wrong* if the system has selected one of the wrong answers;
- *NoA* if the system has decided not to answer the question. Where the system returned a hypothetical answer, this answer was assessed as *NoA_R* in the case of it being correct or *NoA_W* if it was wrong.

After previous years' experience, we realized that advancing the state of the art requires systems ability to decide whether all candidate answers were incorrect or not. In this way, systems able to take this decision should be rewarded over systems that just rank answers.

This is why we introduce in 2013 an explicit assessment focus on testing the ability to reject candidate answers when they are incorrect. We implemented this change by introducing in our tests a portion of questions (39%) where none of the options are correct and including a new last option in all questions: "None of the above answers is correct" (NCA).

It is important to remark that a *NoA* answer is different to a "None of the answers above is correct" (NCA) answer. The former means that the system does **not return any candidate answer** because it is not confident about giving the correct answer, while the latter means that the system rejects the other candidate answers **but returns a response** that will be assessed as *Right* or *Wrong*.

Participant systems were evaluated from two different perspectives:

1. A question-answering approach, as in the traditional QA evaluation campaigns, where we just evaluate the ability of systems answering a set of questions and rank systems according to the final value given by a measure.
2. A reading-test evaluation, obtaining figures for each particular reading test and topics. This perspective permits us to evaluate whether a system was able to understand a document and to what degree. More in detail, we evaluate if the system is able to pass each test, in a similar way to humans with RC tests, what requires obtaining more than 0.5 of *c@1*. This is a kind of evaluation studied with more detail in the pilot Entrance Exams task.

5.1 Evaluation Measure

c@1 has been the main evaluation in all the tasks celebrated in this Lab. *c@1* was firstly introduced in ResPubliQA 2009 [9] and is fully described in [5]. The formulation of *c@1* is given in Formula (1).

$$c@1 = \frac{1}{n}(n_R + n_U \frac{n_R}{n}) \quad (1)$$

where

n_R: number of questions correctly answered.
n_U: number of questions unanswered.
n: total number of questions

The main feature of *c@1* is its consideration of unanswered questions. *c@1* acknowledges unanswered questions in the proportion that a system answers questions correctly, which is measured using the traditional *accuracy* (the proportion of questions correctly answered). Thus, a higher *accuracy* over answered questions, which might be associated to a better validation, would give more value to unanswered questions, and therefore, a higher final *c@1* value. By selecting this measure we wanted to encourage the development of systems able to check the correctness of their responses because NoA answers add value to the final value, while incorrect answers do not.

As a secondary measure, we also provided scores according to *accuracy* (see Formula (2)), the traditional measure applied to past QA evaluations at CLEF. We define a*ccuracy* considering both answered and unanswered questions.

$$accuracy = \frac{n_R + n_{UR}}{n} \tag{2}$$

where

n_R: number of questions correctly answered.
n_{UR}: number of unanswered questions whose candidate answer was correct.
n: total number of questions

5.2 Question Answering Perspective Evaluation

The Question Answering perspective is focused on measuring systems' performance over a set of questions without considering the ability of a system to pass tests associated with documents. This is an approach similar to the one applied in QA@CLEF campaigns before 2011.

The information considered for each system at this level is:

- Total number of questions *ANSWERED*. This number is divided into:
  - total number of questions *ANSWERED* with a *RIGHT* answer,
  - total number of questions *ANSWERED* with a *WRONG* answer.

- Total number of questions *UNANSWERED* (a *NoA* response was given). This number is divided into:
  - total number of questions *UNANSWERED* with a *RIGHT* candidate answer,
  - total number of questions *UNANSWERED* with a *WRONG* candidate answer,
  - total number of questions *UNANSWERED* with an *EMPTY* candidate answer.

The following scores are calculated from this information:

- An overall *c@1* score over the whole collection (the set with 160 questions),
- A *c@1* score for each topic (40 questions for each topic),

- An overall *accuracy* score (over the 160 questions of the test collection, considering also the candidate answers given to unanswered questions as it has been explained above),
- The proportion of answers correctly discarded (see Formula (3)) in order to evaluate the validation performance.

$$correctly_{discarded} = \frac{n_{UW} + n_{UE}}{n_{UR} + n_{UW} + n_{UE}} \quad (3)$$

where:

n_{UR}: number of unanswered questions whose candidate answer was correct
n_{UW}: number of unanswered questions whose candidate answer was incorrect
n_{UE}: number of unanswered questions whose candidate answer was empty

5.3 Reading Perspective Evaluation

The objective of the reading perspective evaluation is to offer information about the performance of a system "understanding" the meaning of each single document. This understanding is evaluated by means of the proposed multiple-choice tests. Each system has to pass a test about a given document similar to the evaluation of RC of new language learners, what was explored in more detailed in the Entrance Exams subtask.

The evaluation is performed taking as reference the *c@1* scores achieved for each test (one document with its ten questions). Then, these c@1 scores can be aggregated at topic and global levels in order to obtain the following values:

- Median, average and standard deviation of *c@1* scores at test level, grouped by topic,
- Overall median, average and standard deviation of *c@1* values at test level.

The median *c@1* is provided under the consideration that it can be sometimes more informative at reading level than average values. This is because median is less affected by outliers than average, and therefore it provides more information about the ability of a system to understand a text.

We consider that a system passes a test according to this evaluation perspective if it achieves a score equal or higher than 0.5.

5.4 Random Baseline

This baseline randomly selects an answer from the set of candidate answers. Since there is one correct option among five, the overall result of this random baseline is 0.2 (both for *accuracy* and for *c@1*). Systems applying a reasonable kind of processing and reasoning should be able to outperform this baseline.

5.5 NCA Baseline

The introduction of the "None of the above answers is correct" option in meaningful proportion, a 39% of questions, allows defining a baseline baseline for a dummy system that always returns this option. This baseline obtained a *c@1* of 0.39.

6 Lessons Learned

Reader will find the quantitative evaluation and results of all runs in all tasks inside the Working Notes Overview papers available on-line from CLEF site. Here we enumerate the main conclusions drawn from this experience.

If we look at the average results in the Main Task along the three years (Table 1), they are close to 0.25 (slightly above from random at 0.20). In general, individual systems select an incorrect answer over the correct one in most cases. There is one notable exception, a system able to give more correct answers than incorrect ones, achieving in each edition a value than 0.5 of c@1.

Table 1. Overview of results 2011-2013 Main Task

	2011	**2012**	**2013**
Average c@1	0.21	0.26	0.24
Best c@1	0.57	0.65	0.59
Average % of unanswered questions	38%	17%	9%

Table 1 shows also how the percentage of unanswered questions decreased in each edition, despite the fact that c@1 values remain similar. This means that systems decision about answering a question or leaving it unanswered had little improvement. Therefore, it seems systems are increasing the risk of giving incorrect answers instead of focusing on developing better validation technologies, as it was expected with the proposal of this task. Possibly, the evaluation measure is not penalizing enough the increase in number of incorrect answers.

This reflection links with the main conclusion of Entrance Exams 2013. Entrance Exams is a very difficult scenario, even challenging for humans. Thus, we can learn from the strategies humans follow to select the correct answer. In most cases, the only way to determine the correct option is by discarding the rest of candidate answers. In other words, *there is more value on developing strategies to discard incorrect answers than strategies to select correct ones.*

Coming back to the Main Task 2013, the correct option was NCA ("none of the above is correct") for 39% of questions. This baseline beats all systems except one, and would have been a good starting point to develop a strategy that decides more carefully on giving an answer only when there is evidence enough.

During the three years of the evaluation, the methodology received several refinements, trying to assess better the level of system performance in deep understanding. One key novelty was the introduction in 2013 of auxiliary questions, reformulating some main questions by reducing the need for inference. This innovation clearly

illustrated which types of reasoning systems were better or worse at. We discover systems find difficulties in questions requiring to connect facts as for example in "Who is the wife of the first president of X?" instead of "Who is the wife of Y?"

Another lesson learned is that most participants reduced the concept of answer validation simply to the task of answer ranking. For this purpose, they develop similarity based approaches that do not decide whether there is a correct answer or not among candidates. Generally, they simply trust the ranking score to exceed a given threshold. So, returning to the question of whether systems achieved enough performance to ensure that there will be a qualitative difference when trying full QA scenario, the answer is: possibly not.

Over the years, it has become clear that groups working on Question Answering are not making use of background knowledge collections very much. At most, systems might locate some possibly relevant material from the background collection through simple matching, and then use associated information to help rank the potential answers. Tying in with the point above on answer ranking above, it indicates the difficulty to introduce inference/reasoning into processing.

Regarding the construction of background collections, we learned it is very difficult to adequately define Background Knowledge, and to specify the types and sources that must be considered to solve the full QA scenario. There are increasingly more sources of linked / relational data that, potentially, can be used. However, language goes beyond a predefined set of relations among entities and values. That was the reason to propose the use of text collections inviting participants to acquire propositional knowledge useful for textual inferences. We have not obtained much of value in this regard.

Despite the difficulty on defining Background Knowledge, we have learned that if we want to use text collections to contextualize system readings, we must be very careful to not introduce any kind of bias. So, the idea of creating a background collection able to contextualize a single text can be formulated as a classical Information Retrieval task: retrieve all relevant documents and only them. Any methodological approach must take this ideal as reference and try to approximate it as much as possible.

We believe that the resources generated so far by QA4MRE will serve to measure progress in this direction, since they form a 100% reusable benchmark in several languages.

7 Related Work

Over the last years, the QA Track at CLEF has changed its evaluation methodology in order to promote deeper text understanding. Clearly, the task of retrieving just text excerpts (facts, sentences, paragraphs, or documents) is not enough to develop the technology. Besides QA, other evaluation activities were also performed which required deeper analyses of texts, for example Recognizing Textual Entailment (RTE), Answer Validation (AV), and Knowledge Base Population (KBP).

Question Answering: a system receives questions formulated in natural language and returns one or more exact answers to these questions, possibly with the locations from which the answers were drawn as justification. The evaluation of QA systems began at the Text Retrieval Conference (TREC) and was continued at the Cross Language Evaluation Forum (CLEF) in the EU and at the NII-NACSIS Test Collection for IR Systems (NTCIR) in Japan. Most of the questions used in these evaluations ask about facts (e.g., Who is the president of XYZ?) or definitions (e.g., What does XYZ mean?). Since systems could search for answers among several documents (using IR engines), it was generally possible to find in some document a 'system-friendly' statement that contained exactly the answer information stated in an easily matched form. This made QA both shallow and relatively easy.

Recognizing of Textual Entailment (RTE): a system must decide whether the meaning of a text (the Text T) entails the meaning of another text (the Hypothesis H): whether the meaning of the hypothesis can be inferred from the meaning of the text [14]. RTE systems have been evaluated at the RTE Challenges, whose first competition was proposed in 2005. The RTE Challenges encourage the development of systems that have to treat different semantic phenomena.

Answer Validation Exercise (AVE) [6, 15, 16]. A combination of QA and RTE evaluations, Answer Validation (AV) is the task of deciding, given a question and an answer from a QA system, whether the answer is correct or not. AVE was a task focused on the evaluation of AV systems and it was defined as a problem of RTE in order to promote a deeper analysis in QA.

Another application of RTE, similar to AVE, in the context of Information Extraction was performed in a pilot task at the RTE-6 with the aim of studying the impact of RTE systems in Knowledge Base Population (KBP). The objective of this pilot task is to validate the output of participant systems at the KBP slot-filling task that was included in the Text Analysis Conference (TAC). Systems participating at the KBP slot-filling task must extract from documents some values for a set of attributes of a certain entity. Given the output of participant systems at KBP, the RTE KBP validation pilot consists of deciding whether each of the values detected for an entity is correct according to the supporting document. For taking this decision, participant systems at the RTE KBP validation pilot receive a set of T-H pairs, where the hypothesis is built combining an entity, an attribute and a value.

Other efforts closer to our proposal for evaluating systems understanding took place as the ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as evaluation for computer-based language understanding systems. This workshop proposed to evaluate understanding systems by means of Reading Comprehension (RC) tests. The evaluation consisted of a set of texts and a series of questions about each text. Quite interestingly, most of the approaches presented at that workshop showed how to adapt QA systems to such kind of evaluation.

A more complete evaluation methodology of MR systems has been reported in [11], where the authors also proposed to use RC tests. However, the objective of these tests was to extract correct answers from documents, which is similar to QA without an IR engine.

8 Conclusions

QA4MRE is characterised by two major innovations. First, there was a transition from traditional Question Answering based on shallow text 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.

This new evaluation was well received by the QA community. Significant lessons were learned from it.

Acknowledgements. Anselmo Peñas and Álvaro Rodrigo have been partially supported by the Research Network MA2VICMR (S2009/TIC-1542) and READERS project (CHIST-ERA). Eduard Hovy was supported by two DARPA grants in Machine Reading.

References

1. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. AI Magazine 31(3) (2010)
2. Morante, R., Daelemans, W.: Annotating Modality and Negation for a Machine Reading Evaluation. CLEF 2011 Labs and Workshop Notebook Papers (2011)
3. Morante, R., Daelemans, W.: Annotating Modality and Negation for a Machine Reading Evaluation. CLEF 2012 Evaluation Labs and Workshop Online Working Notes (2012)
4. Morante, R., Krallinger, M., Valencia, A., Daelemans, W.: Machine Reading of Biomedical Texts about Alzheimer's Disease. CLEF 2012 Evaluation Labs and Workshop Online Working Notes (2012)
5. Peñas, A., Rodrigo, Á.: A Simple Measure to Assess Non-response. In: Proceedings of 49th Annual Meeting of the Association for Compu-tational Linguistics-Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA (2011)

6. Peñas, A., Rodrigo, Á., Sama, V., Verdejo, F.: Overview of the Answer Validation Exercise 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 257–264. Springer, Heidelberg (2007)
7. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Forascu, C., Sporleder, C.: Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. Working Notes, CLEF 2011 (2011)
8. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. Working Notes, CLEF 2012 (2012)
9. Peñas, A., et al.: Overview of resPubliQA 2009: Question answering evaluation over european legislation. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 174–196. Springer, Heidelberg (2010)
10. Sutcliffe, R., Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE Main Task at CLEF 2013. Working Notes, CLEF 2013 (2013)
11. Wellner, B., Ferro, L., Greiff, W., Hirschman, L.: Reading Comprehension Tests for Computer-based Understanding Evaluation. Natural Language Engineering 12(4), 305–334 (2006)
12. Morante, R., Krallinger, M., Valencia, A., Daelemans, W.: Machine Reading of Biomedical Texts about Alzheimer's Disease 2013. Working Notes, CLEF 2013 (2013)
13. Peñas, A., Miyao, Y., Hovy, E., Forner, P., Kando, N.: Overview of QA4MRE 2013 Entrance Exams Task. Working Notes, CLEF 2013 (2013)
14. Dagan, I., Glickman, O., Magnini, B.: The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F. (eds.) MLCW 2005. LNCS (LNAI), vol. 3944, pp. 177–190. Springer, Heidelberg (2006)
15. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 237–248. Springer, Heidelberg (2008)
16. Rodrigo, Á., Peñas, A., Verdejo, F.: Overview of the Answer Validation Exercise 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 296–313. Springer, Heidelberg (2009)